Universidade do Minho
Escola de Engenharia

Tiago Emanuel Senra da Cruz

Data Analytics para Variedade de Dados

Outubro 2017

Data Analytics para Variedade de Dados
Tiago Emanuel Senra da Cruz
UMinho | 2017

Tiago Emanuel Senra da Cruz

Data Analytics para Variedade de dados

Dissertação de Mestrado

Mestrado integrado em Engenharia e Gestão de Sistemas de Informação

Trabalho efetuado sob a orientação do

Professor Doutor Jorge Oliveira e Sá

Outubro 2017

DECLARAÇÃO

Nome: Tiago Emanuel Senra da Cruz

Endereço eletrónico: tiago.cruz.1992@gmail.com

Telefone: 912288122

Número do Bilhete de Identidade: 14086235

Título: Data Analytics para Variedade de Dados

Orientador(es): Professor Jorge Oliveira e Sá

Ano de conclusão: 2017

Designação do Mestrado: Mestrado integrado em Engenharia e Gestão de Sistemas de Informação

Universidade do Minho, __/__/____

Assinatura: _____

Agradecimentos

Aos meus pais, agradeço por todo apoio quer parental como financeiro, pois sem eles não tinha condições para chegar a este patamar. A eles devo tudo pois nunca desistiram de mim nem dos meus sonhos, e puseram sempre os filhos como primeira prioridade. São sem dúvida um enorme exemplo para mim e é a eles que dedico em primeira mão este objetivo cumprido. Ao meu irmão, que acompanhou todo o meu percurso e esteve sempre lá quando precisei de ajuda.

Agradeço aos meus avós (os que estão cá e os que já partiram), aos meus tios, aos meus primos, aos meus padrinhos, a toda a família, por toda a motivação e preocupação de todo o meu percurso académico, sem dúvida não poderia ter uma família mais unida que esta.

Agradeço ao meu orientador, professor Doutor Jorge Oliveira e Sá, por todo o apoio, paciência, e disponibilidade durante esta longa jornada de um ano. Foi uma peça crucial para me guiar no caminho certo no decorrer desta dissertação.

A todos os meus amigos, agradeço por estarem sempre presentes quando precisei, pois, a vossa companhia não tem preço.

À Sónia, o meu amor, por estar comigo deste o início desta etapa, foste o meu suporte e a minha força. Estiveste sempre lá quando precisei de ti, e deste-me sempre motivação para ultrapassar todos os meus obstáculos. Obrigado por tudo!

A todos aqueles que passaram a meu lado neste percurso académico, quer professores como amigos, por me ajudarem e ensinarem tudo o que sei hoje, pois vocês fizeram-me crescer como pessoa, e não apenas intelectualmente. Obrigado a todos!

Resumo

A *Internet* permitiu que os gestores das organizações tivessem acesso a grandes quantidades de dados, e esses dados são apresentados em diferentes formatos, em concreto estruturados, semiestruturados e não estruturados. Neste momento, esta variedade de dados é, em parte, oriunda das redes sociais, onde os utilizadores geram conteúdos diversificados como por exemplo *websites*, blogs, imagens, vídeos entre outros, mas não só, mas também as máquinas são capazes de partilhar informações entre si, ou máquinas com pessoas, através da *internet*. Os tipos de dados gerados já não são apenas do formato estruturado, mas semiestruturados e não estruturados.

Face à variedade de dados disponíveis, é realçada a importância de analisar estes dados para que os gestores possam tirar partido deles para a tomada de decisão. Verifica-se para os dados estruturados que já existem técnicas validadas, estudadas e maduras, mas para os outros tipos de dados semiestruturados e não estruturados tal não se verifica.

O objetivo desta dissertação passou por perceber face à variedade de dados existente, nomeadamente dados não estruturados e semiestruturados, que tipo de informações é possível retirar desses dados, através da sua análise.

Foi realizada uma experiência, utilizando um *dataset* com comentários de carros e um conjunto de imagens para ilustrar o modelo do carro. Foram utilizadas quatro técnicas de análise distintas, sendo elas: Processamento da Linguagem Natural; Análise de Sentimento; Análise de Emoção e Reconhecimento de imagens; para retirar informações desses dados. De seguida foi procedido a criação de uma plataforma analítica e a sua visualização através de *dashboards*. Verifica-se que é assim possível retirar um conjunto de informações, como a análise de sentimento, emoção, quais as componentes que as pessoas gostam mais/menos de um determinado carro, ou sobre uma categoria de carro, entre outras informações relevantes.

Palavras-chave: Análise de Dados, Tipos de Dados, Técnicas de análise de dados

Abstract

The Internet has made it possible for organizations managers to have access to substantial amounts of data, and these data are presented in different formats, namely structured, semi-structured and unstructured. Now, this data variety is party derived from social networks, where users generate diverse content such as websites, blogs, images, videos, among others, but not only, also machines are able to share information between themselves, or machines with people, through the internet. The data formats generated are no longer just the structured, but also semi-structured and unstructured.

Given the data variety available, the importance of analyzing this data is emphasized so that organization managers can benefit from it for decision-making. For structured data, there are already studied, validated and mature techniques, but for the other formats this is not the case.

The purpose of this dissertation was to perceive due the data variety available, namely semi-structured and unstructured data, which kind of information can be extracted from these data, through its analysis.

An experiment was conducted using a dataset containing car reviews and a set of images to illustrate the car model. Four different analysis techniques were used to extract information: Natural Language Processing; Sentiment analysis; Emotion analysis; and Image recognition; from these data. Then the next step was the creation of an analytical platform and is visualization through dashboards. It turns out that it is possible to withdraw a set of information, such as the feeling analysis, emotion, witch components people like the most/ less of a car, or about a car category, among other relevant information.

Keywords: Data Analytics, Analytics, Data Types, Analytic Techniques

Índice

Índice de Figuras

Índice de Tabelas

viii

Siglas e Acrónimos

AI: Artificial Intelligence

API: Application programming Interface

BI&A: Business Intelligence and Analytics

BPM: Business Process Management

BSON: Binnary JSON

CRFs: Conditional Random Fields

CRM: Customer Relationship Management

ERP: Enterprise Resource Planning

ETL: Extract Transform and Load

FSD: First Story Detection

GTE: Generalized Transfer Entropy

IDE: Integrated Development Environment

IJBDI: International Journal of Big Data Intelligence

IoT: Internet of Things

JSON: JavaScript Object Notation

KRR: Kernel Ridge Regression

LDA: Latent Dirichlet Allocation

MIL: Multi Instance Learning

NLP: Natural Language Processing

OEM: Object Exchange Model

OLAP: Online Analytical Processing

OWL: Web Ontology Language

PLSV: Probabilistic Latent Semantic Visualization

RDF: Resource Description Framework

RFID: Radio Frequency Identification

SaaS: Software as a Service;

SGBD: Sistema de Gestão de Base de Dados

SVM: Support Vector Machine

TEM: Probabilistic Three-way Entity Model

WSDM: International Conference on Web Search and Data Mining

XML: eXtensible Markup Language

YAML: YAML Ain't Markup Language

1. Introdução

Neste capítulo, é apresentado o enquadramento e motivação desta dissertação, a finalidade e principais objetivos, e por fim a estrutura do relatório.

1.1. Enquadramento e Motivação

Com o aumento da utilização da *internet* por parte das pessoas e organizações, a quantidade de informação cresceu exponencialmente. O universo digital apresenta uma diversidade de dados, como dados de redes sociais, sensores *RFID*, dados geográficos, dados de blogs e *websites*, dados provenientes de dispositivos móveis, entre outros. Esta diversidade de dados é acompanhada por uma variedade de tipo dados e consiste numa característica a ter em consideração pois a sua análise pode trazer informações interessantes para as organizações (Russom, 2011). De acordo com Cukier (2010), apenas 5% dos dados são do formato estruturado, encontrados usualmente em base de dados, os restantes 95% são do formato semiestruturado ou não estruturado e usualmente não são explorados. Isto revela que existe um grande volume de dados que deve ser analisado de forma a se perceber o seu potencial. A análise dos tipos de dados não estruturado e semiestruturado pode realmente fornecer informações úteis a uma organização, caso lhes seja dada a devida atenção (algo que não acontece em grande parte das organizações).

A análise desta variedade de dados pode ser um aspeto importante para as organizações, pois pode oferecer informações pertinentes aos gestores de uma organização, facilitando o processo de tomada de decisões.

Neste momento, existem diferentes técnicas de análise de dados capazes de retirar informações para análise, de forma a que os gestores possam tirar partido deles na tomada de decisão. Considera-se que as técnicas de análise de dados do tipo estruturado já estão maduras e validadas pela comunidade científica, mas tal não se verifica para os dados do tipo semiestruturado e não estruturado.

1.2 Objetivos e Resultados Esperados

A finalidade da presente dissertação passa por estudar para os diferentes tipos de dados que técnicas de análise de dados existem, com principal foco para as técnicas de análise orientadas para

dados não estruturados e semiestruturados, assim como perceber de que forma é possível retirar informações relevantes da variedade de dados existente.

Assim, espera-se elaborar uma prova de conceito, que a partir de um conjunto de dados de teste, serão aplicadas técnicas de análise de forma a se obter informações sobre esses dados, com o fim de se proceder a sua análise.

Desta forma, para atingir o desiderato desta dissertação é necessário traçar um conjunto de objetivos.

Os objetivos propostos são os seguintes:

- descrever dados e tipos de dados;
- levantar técnicas de análise de dados para dados semiestruturados e não estruturados, e descrever os tipos de análise de dados;
- descrever técnicas de análise de dados existentes na bibliografia;
- identificar ferramentas de análise de dados; e
- efetuar uma prova de conceito, onde serão aplicadas técnicas de análise aos dados e ferramentas identificadas no objetivo anterior.

1.3 Abordagem Metodológica

Pretende-se com esta secção, apresentar a questão de investigação, a metodologia utilizada, e por fim a aplicação da metodologia neste documento.

1.3.1 Questão de Investigação

A questão de investigação consiste no problema que se pretende resolver/estudar. A questão consiste no seguinte:

"Tirando partido da variedade de dados existente, nomeadamente dados não estruturados e semiestruturados, será possível retirar informações pertinentes desses dados, através da sua análise?"

1.3.2 Metodologia de Investigação

A abordagem metodológica adotada nesta dissertação foi a *Design Science Research for Information Systems* (Peffers, Tuunanen, Rothenberger, & Chatterjee, 2008) pois esta refere-se à agregação de conhecimento adquirido através de uma revisão de literatura e conceitos teóricos relevantes, com o *design* de soluções para dar resposta a problemas do mundo real, ver figura 1.

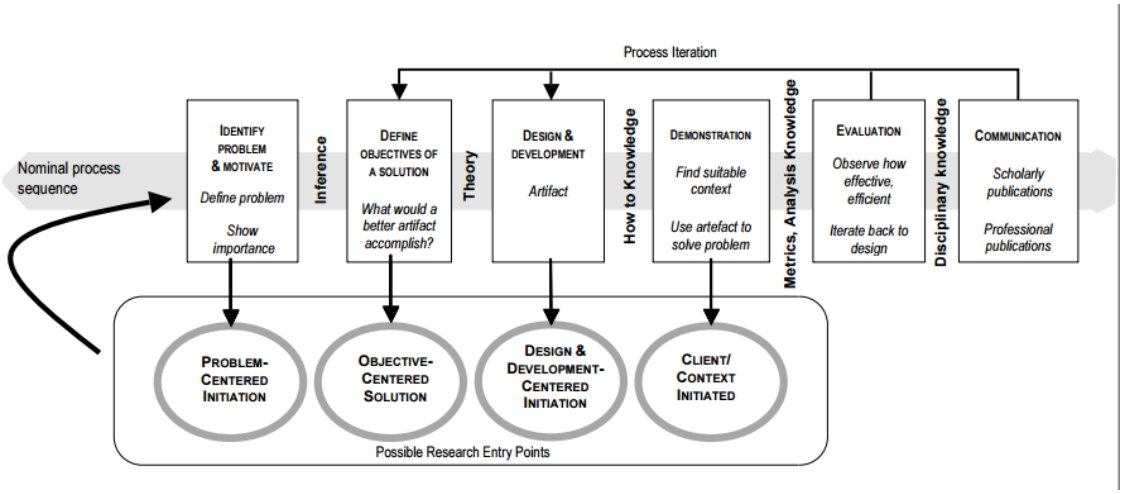


Figura 1 - Design Science Research Methodology (DSRM) Process Model.

(Peffers, Tuunanen, Rothenberger, & Chatterjee, 2008)

O primeiro passo consiste na identificação do problema e motivação. Como a definição do problema será utilizada para desenvolver um artefacto que pode efetivamente fornecer uma solução, pode ser útil ir ao detalhe do problema conceitualmente para que a solução possa capturar a sua complexidade. Justificar o valor de uma solução implica duas coisas: motivar o investigador e o público da pesquisa a irem ao encontro da solução e a aceitar os resultados, e ajudar a entender o raciocínio associado à compreensão do investigador sobre o problema. Os recursos necessários para esta atividade incluem o conhecimento do estado da arte do problema e a importância da sua solução.

O segundo passo passa por definir os objetivos para a solução. Os objetivos podem ser quantitativos, onde através de indicadores é possível medir o desempenho de uma solução; ou

qualitativos, momento onde o principal objetivo consiste em interpretar um determinado fenômeno, através da sua observação, descrição, compreensão e o significado do seu comportamento. Os objetivos devem ser inferidos racionalmente a partir da especificação do problema. Os recursos necessários para esta atividade incluem o conhecimento do estado dos problemas e soluções atuais, se existirem, e a sua eficácia.

O terceiro passo refere-se à fase de design e desenvolvimento. Consiste na criação do artefacto. Este artefacto pode assumir várias formas: modelos, métodos, instâncias (Von Alan, March, Park, & Ram, 2004) "novas propriedades de recursos informacionais e/ou sociais" (Järvinen, 2007). Conceptualmente, um artefacto de pesquisa de design pode ser qualquer objeto projetado no qual uma contribuição de pesquisa está incorporada no projeto. Esta atividade inclui a determinação da funcionalidade desejada do artefacto e a sua arquitetura, e de seguida proceder à criação do artefacto real. Os recursos necessários para passar de objetivos de design ao desenvolvimento, incluem o conhecimento da teoria que pode servir como base para uma possível solução.

O quarto passo refere-se à demonstração. Demonstrar o uso do artefacto para resolver uma ou mais instâncias do problema. Este passo pode envolver o seu uso na experimentação, simulação, estudo de caso, prova, ou outra atividade apropriada. Os recursos necessários para a demonstração incluem o conhecimento efetivo de como usar o artefacto para resolver o problema.

O quinto passo passa pela Avaliação. Observar e medir como o artefacto suporta uma solução do problema definido. Esta atividade envolve a comparação dos objetivos de uma solução com os resultados efetivamente adquiridos da utilização do artefacto na demonstração. Este passo requer conhecimento de métricas e técnicas de análise relevantes. Dependendo da natureza do problema e do artefacto, a avaliação pode assumir diversas formas. Poderia incluir itens para servir de comparação da funcionalidade do artefacto com os objetivos da solução, medidas de desempenho quantitativo, como orçamentos ou itens produzidos, resultados de pesquisas de satisfação, *feedback* de clientes ou simulações. Poderia incluir medidas quantitativas de desempenho do sistema, como o tempo de resposta ou disponibilidade. Conceptualmente, tal avaliação poderia incluir qualquer tipo de prova empírica apropriada ou prova lógica. No final desta atividade, os investigadores podem decidir iterar novamente para o passo três e tentar melhorar a eficácia do artefacto, ou para continuar a comunicação e deixar melhorias futuras para projetos subsequentes. A natureza do ponto de investigação pode determinar se tal iteração é viável ou não.

O sexto e último passo consiste na comunicação. Comunicar o problema e a sua importância, o artefacto, a sua utilidade e inovação, o rigor do seu design e a sua efetividade para investigadores e outros públicos relevantes, como profissionais na área, quando apropriado. Em pesquisas académicas, os investigadores podem usar a estrutura desse processo para estruturar um artigo, assim como a estrutura nominal de um processo de pesquisa empírica (definição do problema, revisão de literatura, desenvolvimento de hipóteses, coleta de dados, análise, resultados, discussão e conclusão) consiste numa estrutura comum para trabalhos empíricos de investigação. A comunicação requer conhecimento da cultura disciplinar (Peffers, Tuunanen, Rothenberger, & Chatterjee, 2008).

1.3.3 Aplicação da Metodologia

De seguida, irá ser explicado a forma como a metodologia *Design Science Research for Information Systems* ajudou na realização do trabalho de investigação descrito neste documento.

A primeira fase consistiu na definição do problema e motivação definidos no ponto 1.1.

Na segunda fase, pretende-se definir um conjunto de objetivos para a realização da dissertação. Este conjunto de objetivos vão servir para ajudar a responder ao problema que foi proposto. Está proposto um conjunto de objetivos identificados no ponto 1.2.

A terceira fase consiste no design e desenvolvimento do artefacto. Pretende-se provar que é possível retirar informações pertinentes de dados não estruturados e semiestruturados, utilizando diferentes técnicas de análise de dados, e ter como *output* um conjunto de informações relevantes para análise, ilustrado no ponto 3.

A quarta e quinta fase, passam por demonstrar e avaliar a performance do artefacto para responder ao problema proposto. Pretende-se mostrar que ao inserir dados de diferentes formatos, o artefacto é capaz de retirar informações úteis para análise, ilustrado no ponto 3.6. Para ilustrar que os dados são úteis para análise, está previsto que o artefacto seja complementado com uma plataforma analítica de forma a facilitar a análise dos dados obtidos, presente no ponto 3.7.

A última fase, refere-se à comunicação. Nesta fase, espera-se a apresentação dos principais resultados obtidos do estudo deste problema. Pretende-se apresentar a importância da construção deste documento, a sua utilidade, e o contributo que proporciona para a comunidade científica.

Serão apresentados os resultados obtidos, e também mostrar que se consegue dar resposta à questão de investigação. Também está previsto a divulgação dos resultados numa ou mais conferências científicas.

1.4 Organização do Documento

No capítulo 1 são apresentados o enquadramento e motivação, objetivos e resultados esperados, questão de investigação, descrição da metodologia de investigação utilizada e a sua aplicação neste trabalho de dissertação.

No capítulo 2, referente à Revisão de Literatura é apresentado um conjunto de informações que serviram de apoio teórico para a realização deste documento.

No capítulo 3, passa pela apresentação da construção do artefacto, contendo um conjunto de subpontos: técnicas de análise utilizadas, apresentação dos *datasets* e ferramentas utilizadas; descrição da arquitetura global do artefacto; descrição de como foi realizada a extração de dados; teste do artefacto; e por fim, construção da plataforma analítica.

No capítulo 4, é apresentada a discussão dos resultados, através da análise dos dados através da plataforma analítica.

No capítulo 5, são apresentadas as conclusões, trabalho futuro e limitações da realização deste documento.

2 Enquadramento Conceitual

Neste capítulo são apresentados os conceitos teóricos relevantes para ajudar na resolução do problema proposto, passando pela identificação do conceito de dados, tipo de dados, análise de dados, tipos de análise de dados, técnicas de análise de dados e ferramentas analíticas.

2.1 Dados

Nesta secção irá ser ilustrado a definição de dados, seguido da descrição dos tipos de dados.

2.1.1 Definição de Dados

Neste ponto, é apresentado um conjunto de definições de dados sobre as perspetivas de diferentes autores.

Dados consistem num conjunto de valores qualitativos ou quantitativos de objetos e eventos. Os dados são medidos, recolhidos, analisados, e podem ser visualizados utilizando gráficos, imagens ou outras ferramentas de análise (Ackoff, 1989).

Os dados são variáveis, simplesmente existem e não possuem significado além da sua existência ou podem existir em qualquer forma, podendo ser utilizáveis ou não (Bellinger, Castro, & Mills, 2004).

Os dados são um conjunto de factos discretos e objetivos sobre eventos (Davenport & Prusak, 1998). Os dados descrevem apenas uma parte do que aconteceu e não fornecem julgamento ou interpretação e nenhuma base sustentável de ação. Os dados não dizem nada sobre a sua importância e relevância.

2.1.2 Tipos de Dados

Os dados podem ser classificados da seguinte forma:

- **Dados Estruturados**: consistem em dados organizados numa estrutura predefinida, como uma base de dados relacional, de forma a que os seus elementos possam ser endereçados para um processamento e análise mais eficaz. Os dados estruturados possuem a vantagem de ser facilmente inseridos, armazenados, analisados, e consultados através de *queries SQL* (do inglês *Structured Query Language*). Todos estes

dados são guardados num sistema de gestão de base de dados (SGBD) e podem ser utilizados em sistemas como por exemplo:

- ***CRM*** (do inglês *Customer Relationship Management*): consiste numa abordagem que coloca o cliente como principal foco dos processos de negócio, com o objetivo de perceber e antecipar as suas necessidades, de forma a melhorar o contacto entre a organização e o cliente;
- ***Data Warehouse***: refere-se a um repositório de dados com o objetivo de armazenar informações detalhadas relativamente a uma organização que possibilita uma análise de grandes volumes de dados, que podem ser recolhidos de diferentes fontes para dar suporte à tomada de decisão das organizações; e
- ***ERP*** (do inglês *Enterprise Resource Planning)*: consiste num conjunto de sistemas de informação que integram todos os dados de uma organização num único sistema. Apresenta como vantagens a otimização do processo de tomada de decisão, a eliminação de redundância de atividades, a redução do tempo de execução dos processos de gestão, etc., entre outros (Kimball & Ross, 2011).

- Dados semiestruturados: consistem numa coleção de dados heterogéneos, que não possuem uma estrutura rígida (Asai, et al., 2004). Possuem como característica uma estrutura irregular, isto é, uma estrutura mais descritiva ao contrário dos dados estruturados que apresentam uma estrutura prescritiva. Para além disso, os dados semiestruturados podem não possuir um formato predefinido ao contrário dos dados estruturados que possuem sempre um formato predefinido. Como exemplo de dados semiestruturados, temos:
  - ***OEM*** (do inglês *Object Exchange Model*): modelo criado para realizar trocas de dados semiestruturados entre base de dados orientadas a objetos;
  - ***XML*** (do inglês *eXtensible Markup Language)*: tem como principal propósito facilitar a troca de informações através da *internet* utilizando uma linguagem padronizada de marcação genérica;
  - ***RDF*** (do inglês *Resource Description Framework)*: consiste numa linguagem para representar informação na *internet* e são modelos ou fontes de dados com o principal objetivo de criar um modelo simples de dados, i.e., com uma semântica formal; e

- *OWL* (do inglês *Web Ontology Language*): é uma linguagem para definir e instanciar ontologias na *Web*, uma ontologia *OWL* pode incluir descrições de classes e as suas respetivas propriedades e os seus relacionamentos.

- Dados não estruturados: são dados que não possuem uma estrutura pré-definida ou que não estão organizados de uma forma pré-definida. São dados muito imprevisíveis e não seguem regras. Usualmente consistem numa grande quantidade de informação que, até recentemente, era desprezada pelos sistemas de *Business Intelligence*. Este conjunto de dados está a ser alvo de grande estudo recentemente devido ao avanço na área de *Internet of Things* (IoT), que permitiu que grandes quantidades de dados, a maior parte com formato não estruturados, fossem acessíveis de forma a que as organizações consigam tirar proveito para os seus negócios (Baars & Kemper, 2008). Este conjunto de dados pode ser oriundo de diferentes fontes como sensores, ficheiros de áudio, imagens, dados de redes sociais entre outros.

2.2 Análise de Dados

Neste capítulo, é apresentado uma perspetiva histórica sobre a análise de dados até ao momento.

Business intelligence & Analytic (*BI&A*) consiste numa abordagem centrada em dados, com na gestão de bases de dados históricos. Existem três momentos de *BI&A* que serão descritos neste ponto e que são identificados como *BI&A 1.0, BI&A 2.0 e BI&A 3.0* (Chen, Chiang, & Storey, 2012). Esta abordagem, baseia-se fortemente em diversas tecnologias de aquisição, extração e análise de dados (Chaudhuri, Dayal, & Narasayya, 2011; Turban, Sharda, Aronson, & King, 2008; Watson & Wixon,2007).

As tecnologias e aplicações de *BI&A* que são adotadas por profissionais podem ser consideradas como *BI&A 1.0*, onde os dados são maioritariamente estruturados, adquiridos pelas organizações através de vários sistemas legados, e muitas vezes armazenados em sistemas de gestão de base de dados. As técnicas analíticas usadas normalmente nestes sistemas, baseiam-se principalmente em métodos estatísticos desenvolvidos na década de 1970 e nas técnicas de prospeção de dados desenvolvidas na década de 1980.

A gestão de dados e *data warehouse* (Kimball & Ross, 2011) são considerados os fundadores de *BI&A*. Design de *data marts* e ferramentas para Extração, Transformação e Carregamento (*ETL*

do termo em inglês *Transform, Loading and Extraction*) são essenciais para converter e integrar dados específicos da organização. *Queries* a base de dados, processamento analítico *online* (*OLAP* do termo inglês *Online Analytical Processing*), e ferramentas de *reporting* baseado em gráficos intuitivos são utilizados para explorar características de dados relevantes. A gestão de processos de negócio (*BPM* do termo inglês *Business Process Management*) em conjunto com *scorecards* e *dashboards* ajudam a analisar e visualizar uma variedade de métricas de desempenho.

Existem oito técnicas/modelos que são considerados *BI&A 1.0*: *reporting*, *dashboards, ad hoc query, search-based BI, OLAP*, visualização interativa, *scorecards*, modelos predição, prospeção de dados (Sallam, Richardson, Hagerty, & Hostmann, 2011).

No início do século XXI, a *internet* e a *web* começaram a oferecer oportunidades na recolha de dados e investigação analítica. A Inteligência *Web*, análise da *web* e os conteúdos gerados pelos utilizadores são recolhidos através de sistemas sociais e sistemas *crowdsourcing* (Bitterer, 2011). Estes inauguraram uma nova era de investigação *BI&A 2.0*, centrado na análise de texto e da *web*, para conteúdos *web* não estruturados. Uma quantidade enorme de dados sobre uma organização, indústria, produtos, clientes pode ser recolhida através da *web*, organizada e visualizada através de técnicas de prospeção de texto e *web*.

A comunidade de investigadores da área de marketing acredita que as redes sociais são uma oportunidade única para as organizações interagirem com o mercado de uma forma muito mais interativa, em vez do tradicional contacto entre as organizações e o cliente (Lusch, Liu, & Chen, 2010). Ao contrário das tecnologias *BI&A 1.0* que usualmente estão integradas no sistema de tecnologias de informação do Departamento Comercial de uma organização, sistemas *BI&A 2.0* irão necessitar de integrar técnicas maduras e escaláveis na prospeção de texto: extração de informação; identificação de tópicos; prospeção de opiniões; assim como prospeção da *web*, análise de redes sociais, e análise espaço-temporal.

Considerando que sistemas *BI&A 2.0* baseados na *web* têm atraído investigadores tanto da academia como profissionais, uma nova oportunidade de investigação em *BI&A 3.0* está a surgir essencialmente devido ao aparecimento da *IoT*, que consiste numa grande quantidade de sistemas diversos como smartphones, tablets, computadores, sensores equipados com *RFID*, entre outros, que através de esquemas de endereçamento exclusivos, são capazes de interagir uns com os outros e cooperar com os seus vizinhos para alcançar objetivos comuns (Giusto, Iera, Morabito, & Atzori, 2010). A maior parte da investigação académica em *BI* móvel ainda se encontra numa fase

embrionária. A década de 2010 promete ser uma era com grande impacto na investigação e desenvolvimento em *BI&A*, tanto para os profissionais como para os académicos (Halper & Krishnan, 2013).

As organizações podem utilizar análises descritivas ou preditivas nos seus projetos. Numa fase inicial, o tipo de ferramentas de análise utilizado depende do problema que a organização pretende resolver. Normalmente, as organizações ainda utilizam apenas um tipo de dados, embora isso possa variar entre organizações nesta fase. Por exemplo, algumas organizações utilizam grandes volumes (mais de 10TB) de dados estruturados armazenados numa ferramenta. A organização pode estar a executar algum tipo de modelo preditivo sobre estes dados. Alternativamente, uma organização (por exemplo, uma editora) pode ter uma gestão aperfeiçoada e utilizar grandes quantidade de conteúdos, mas não é forte na análise. Algumas organizações podem estar a utilizar diferentes tipos de dados, mas não de forma integrada. Por exemplo algumas organizações podem estar primeiramente utilizar dados internos estruturados, mas também a usar dados não estruturados provenientes de redes sociais para outros departamentos da organização.

2.2.1 Tipos de Análise de Dados

As análises de dados podem ser classificadas de diferentes tipos (Evans & Lindner, 2012):

- **Análise Preditiva:** analisar o desempenho passado de forma a prever o futuro, analisando dados históricos, detetando padrões ou relacionamentos nesses dados, e por fim, extrapolar essas relações para a frente no tempo. Por exemplo, um profissional de marketing pretende prever a resposta de diferentes segmentos de clientes numa campanha publicitária; um *trader de commodities* pode desejar prever movimentos de curto prazo nos preços das *commodities;* ou um fabricante de esquis pode querer prever a procura da próxima temporada de esquis, ou uma cor ou tamanho específico. A análise preditiva pode prever o risco de encontrar relacionamentos em dados que não são facilmente identificáveis em análises tradicionais. Utilizando técnicas avançadas, a análise preditiva pode ajudar a detetar padrões ocultos em grandes quantidades de dados de forma a segmentar e agrupar dados em conjuntos coerentes, a fim de prever o comportamento e detetar tendências.

  Análise Preditiva aborda questões como: o que vai acontecer se a procura cair 10% ou se os preços dos fornecedores subirem 5%? Quanto esperamos pagar pelo combustível nos próximos meses? Qual é o risco de perder dinheiro ao investir num novo negócio?

- **Análise Descritiva:** utilizar dados para perceber o desempenho passado e presente de uma organização, e tomar decisões informadas. Estes tipos de técnicas servem para categorizar, caracterizar, consolidar e classificar dados de forma a convertê-los em informações úteis com o propósito de compreender e analisar o desempenho de uma organização. A análise descritiva vai agregar os dados em gráficos e relatórios significativos, por exemplo, sobre orçamentos, vendas, receitas ou custos. Por exemplo, permite que os gestores obtenham relatórios padronizados e personalizados utilizando *queries* de forma a entender o impacto de uma campanha publicitária; analisar o desempenho da organização para encontrar problemas ou oportunidades e identificar padrões e tendências nos dados.

  Análise Descritiva aborda questões como: quanto vendemos em cada região? Qual foi a nossa receita e lucro no último trimestre? Qual foi a quantidade e quais os tipos de reclamações que resolvemos? Qual a fábrica que tem a menor produtividade?

  Análise descritiva também ajuda as organizações a classificar os clientes em segmentos diferentes, o que lhes permite desenvolver campanhas de marketing e estratégias de publicidade especificas.

- **Análise Prescritiva:** utiliza a otimização para identificar as melhores alternativas para minimizar ou maximizar um objetivo, concretamente utiliza técnicas matemáticas e estatísticas de análise preditiva que combinadas com a otimização permitem a tomada de decisões tendo em conta a incerteza nos dados. A análise prescritiva é usada em diversas áreas de negócios, incluindo operações, marketing e finanças. Por exemplo, podemos determinar a melhor estratégia de preços e publicidade para maximizar a receita; a quantidade ideal de dinheiro para armazenar nos ATMs.

  Análise Prescritiva aborda questões como: quanto devemos produzir para maximizar o lucro? Qual a melhor maneira de enviar mercadorias das nossas fábricas para minimizar os custos? Devemos mudar os nossos planos se um desastre natural fechar a fábrica de um fornecedor, e se realmente acontecer, por quanto?

2.2.2 Técnicas de Análise de Dados

De seguida, vai ser apresentado um conjunto de tabelas com técnicas de análise de dados retirados de cinco conferências de investigação sobre análise de dados. Também é apresentado um conjunto de tabelas ilustrando para os diferentes tipos de dados quais as técnicas que se adequam utilizando

as cinco conferências que se apresentam. As tabelas são compostas por um conjunto de pontos, sendo eles: referência, nome do artigo, técnicas de análise, e por fim tipos de dados.

Conferência KDD2014 ano 2014 rank A*

Tabela 1 - Técnicas de análise de dados KDD2014

*ano 2014 rank A**

Referência	Artigo	Técnica/(as) de análise de dados	Tipo de dados
(Sudhof, Goméz, Maas, & Potts, 2014)	Sentiment Expression Conditioned by Affective Transitions and Social Forces	Conditional Random Fields (CRFs) as a Modeling Technique	Não estruturado;
(Dong, et al., 2014)	Knowledge Vault: A Web-Scale Approach to Probabilistic Knowledge Fusion	NLP, HTML trees (DOM)	Não estruturado; Semiestruturado;
(Chen & Liu, 2014)	Mining Topics in Documents: Standing on the Shoulders of Big Data	AMC (Topic Modeling With Automatically Generated Must-Links and Cannot-links)	Não estruturado;
(Mukherjee, Weikum, & Danescu-Niculescu-Mizil, 2014)	People on Drugs: Credibility of User Statements in Health Communities	Subject-Predicate-Object Statement Extraction (NLP), LDA-style Models;	Não estruturado
(Kurashima, Iwata, Takaya, & Sawada, 2014)	Probabilistic Latent Network Visualization: Inferring and Embedding Diffusion Networks	Probabilistic Latent Semantic Visualization (PLSV)	Não estruturado;
(Schubert, Weiler, & Kriegel, 2014)	SigniTrend: Scalable Detection of Emerging Topics in Textual Streams by Hashed Significance Thresholds	First Story Detection (FSD)	Não estruturado;
(Yao, Tong, Xu, & Lu, 2014)	Predicting Long-Term Impact of CQA Posts: A Comprehensive Viewpoint	Kernel Ridge Regression (KRR), Support Vector Regression (SVR)	Estruturado;
(Zhao, Liu, & Cox, 2014)	Safe and Efficient Screening For Sparse Support Vector Machine	Sparse Support Vector Machine	Estruturado/ Não estruturado
(Anagnostopoulos & Triantafillou, 2014)	Scaling Out Big Data Missing Value Imputations: pythia vs. godzilla.	Sequential Multivariate Regression Imputation	Estruturado
(Grabocka, Schilling, Wistuba, & Schmidt-Thieme, 2014)	Learning Time-Series Shapelets	Time-Series Shapelets	Estruturado

Conditional Random Fields (*CRFs*) consiste num método de modelação estatística, muitas vezes aplicado no reconhecimento de padrões e *machine learning*. Os *CRFs* encontram aplicações

em análise superficial, reconhecimento de entidades nomeadas, entre outras tarefas, sendo uma alternativa aos modelos de *Markov*. *CRFs* são frequentemente utilizados para reconhecimento de objetos e segmentação de imagens (Sudhof, Goméz, Maas, & Potts, 2014).

Natural Language Processing refere-se à capacidade de um programa computacional compreender o discurso humano. As técnicas utilizadas foram o reconhecimento de entidades, parte da marcação de discurso, análise de dependência e ligação de entidades (Dong, et al., 2014).

Árvore DOM (do termo inglês *Document Object Model*) consiste numa forma de extrair informação de paginas *web*. Os objetos na árvore *DOM* podem ser endereçados e manipulados através do uso de métodos sobre objetos (Dong, et al., 2014).

Modelo de tópicos consiste num modelo estatístico cuja inferência pode explorar o conhecimento minerado automaticamente, lidar com as questões de conhecimento errado e a transitividade para produzir tópicos superiores, e descobrir tópicos abstratos numa coleção de documentos. Modelo de tópicos é frequentemente utilizado em prospeção de texto para a descoberta de estruturas de semântica no corpo de um texto (Chen & Liu, 2014).

Natural Language Processing, permite extrair ***Subject-Predicate-Object statement*** e é utilizado através da combinação de correspondência de padrões com regras de extração. Isto pode ser realizado de forma superficial, através de sequências de *tokens* de texto, ou em combinação com outras análises linguísticas (Mukherjee, Weikum, & Danescu-Niculescu-Mizil, 2014).

Latent Dirichlet Allocation (*LDA*) é um modelo estatístico que permite um conjunto de observações serem explicadas por grupos não observados que explicam o porquê de partes dos dados serem semelhantes. Modelos do estilo *LDA* têm sido utilizados para sumarizar relatórios de experiências de drogas e para a construção de grandes bases de conhecimento para as ciências da vida e saúde (Mukherjee, Weikum, & Danescu-Niculescu-Mizil, 2014).

Probabilistic Latent Semantic Visualization consiste num modelo probabilístico, no qual extrai tópicos incorporando documentos num espaço de baixa dimensão (Kurashima, Iwata, Takaya, & Sawada, 2014).

First Story Detection tem como foco a identificação do primeiro relatório de um evento, não necessariamente um tópico com grande popularidade (Schubert, Weiler, & Kriegel, 2014).

Kernel Ridge Regression (KRR) e *Support Vector Regression (SVR)* referem-se a uma classe de algoritmos de análise de padrões na área de *machine learning*. A tarefa geral na análise de padrões (como *KRR* e *SVR*) é encontrar e estudar tipos de relações gerais num conjunto de dados, por exemplo: *clusters*; *rankings*; componentes principais; correlações; classificações (Yao, Tong, Xu, & Lu, 2014).

Sparce Support Vector Machine (*SSVM*) consiste num modelo preditivo que consegue remover o ruído e preservar os sinais. Este modelo consegue aprender um caminho como solução utilizando como base um conjunto de parâmetros predefinidos, oferecendo apoio para a seleção do modelo. *SSVM* já foi aplicado com sucesso numa variedade de aplicações de prospeção de dados, prospeção de texto, bioinformática e processamento de imagem (Zhao, Liu, & Cox, 2014).

Sequential Multivariate Regression Imputation estima os valores nulos ajustando uma sequência de modelos de regressão e valores de desenho a partir das distribuições preditivas correspondentes (Anagnostopoulos & Triantafillou, 2014).

Time-Series Shapelets consiste numa técnica de classificação de séries temporais na área de *data mining*, com o objetivo de prever melhor a variável alvo, e tornar o processo mais rápido do que as técnicas normalmente utilizadas do tipo classificação (Grabocka, Schilling, Wistuba, & Schmidt-Thieme, 2014).

Conferência KDD2015 ano 2015 rank A*

Tabela 2 - Técnicas de análise de dados KDD2015

ano 2015 rank A*

Referência	Artigo	Técnica/(as) de análise de dados	Tipo de dados
(Zhou, Liu, & Buttler, 2015)	Integrating Vertex-centric Clustering with Edge-centric Clustering for Meta Path Graph Analysis	Multi-Network Link Prediction	Semiestruturado;
(Nagarajan, et al., 2015)	Predicting Future Scientific Discoveries Based on a Networked Analysis of the Past Literature	Event Extraction	Não estruturado;

Tabela 3 - Técnicas de análise de dados KDD2015 (continuação)

ano 2015 rank A*

Referência	Artigo	Técnica/(as) de análise de dados	Tipo de dados
(Veeriah, Durvasula, & Qi, 2015)	Deep Learning Architecture with Dynamically Programmed Layers for Brain Connectome Prediction	Cross-correlation, Granger Causality, Information Gain, Trained Classifier, Random Score, Generalized Transfer Entropy (GTE), Partial Correlation Statistics	Estruturado;
(Ulanova, et al., 2015)	Efficient Long-Term Degradation Profiling in Time Series for Complex Physical Systems	Time Series Analysis Technique	Estruturado;
(Shashidhar, Pandey, & Aggarwal, 2015)	Spoken English Grading: Machine Learning with Crowd Intelligence	Natural Language Processing Features	Não estruturado;
(Kıcıman & Richardson, 2015)	Towards Decision Support and Goal Achievement: Identifying Action-Outcome Relationships From Social Media	Phrase Segmentation, Canonicalization,	Não estruturado
(Shashidhar, Pandey, & Aggarwal, 2015)	Spoken English Grading: Machine Learning with Crowd Intelligence	Natural Language Processing Features	Não estruturado;
(Kıcıman & Richardson, 2015)	Towards Decision Support and Goal Achievement: Identifying Action-Outcome Relationships From Social Media	Phrase Segmentation, Canonicalization,	Não estruturado
(Kotzias, Denil, De Freitas, & Smyth, 2015)	From Group to Individual Labels using Deep Features	Multi Instance Learning (MIL) Techniques To Text-Based Re-view Data	Estruturado;
(Caballero Barajas & Akella, 2015)	Dynamically Modeling Patient's Health State from Electronic Medical Records: A Time Series Approach	Naive Bayes Classifier; GD-LDA Model	Estruturado; Não estruturado/ Semiestruturado;
(Hayashi, Toyoda, & Kawarabayashi, 2015)	Real-Time Top-R Topic Detection on Twitter with Topic Hijack Filtering	Latent Dirichlet Allocation; Nonnegative Matrix Factorization;	Estruturado e Não estruturado

Multi-Network Link Prediction (*MLI*) explora o *meta-path* para gerar recursos úteis. *MLI* é um *framework* de predição de ligações genéricas e resolve o problema de previsão de ligações multi-rede e as tarefas de previsão de ligação em diferentes redes podem ajudar-se mutuamente (Zhou, Liu, & Buttler, 2015).

Event Extraction consiste numa subcomponente de ***text mining***, com o objetivo de descobrir o que levou ao surgimento de um certo evento, a identificação de associações entre entidades e outras informações num texto (Nagarajan, et al., 2015).

Cross-correlation calcula as pontuações de conectividade com base na correlação cruzada simples sobre dados de ativação neural da série temporal (Veeriah, Durvasula, & Qi, 2015).

Granger Causality consiste num teste de hipótese estatística que calcula uma pontuação baseada na hipótese de a série temporal ser útil na previsão dos outros dados de séries temporais (Veeriah, Durvasula, & Qi, 2015).

Information Gain – Entropy and Gini Index trata todos os pontos dos dados de ativação da série temporal como distribuições independentes e idênticas e calcula uma pontuação de conectividade com base no ganho de informação sobre essas distribuições (Veeriah, Durvasula, & Qi, 2015).

Trained Classifier calcula o coeficiente de correlação parcial com base em algumas características calculadas sobre dados da série de tempo dos neurónios (Veeriah, Durvasula, & Qi, 2015).

Random Score gera uma pontuação aleatória de conectividade. É um método de linha de base simples usado para comparação (Veeriah, Durvasula, & Qi, 2015).

Generalized Transfer Entropy consiste numa estatística não paramétrica que mede a quantidade de transferência direcionada de informações entre dois processos aleatórios. Este é também um dos métodos para prever os conectores dos neurónios cerebrais (Veeriah, Durvasula, & Qi, 2015).

Partial Correlation Statistics consiste em vários filtros de *lower order high pass* e filtros *low pass* aplicados sobre dados de séries temporais. Estes sinais de séries temporais filtrados são utilizados para calcular uma pontuação de conectividade com base no coeficiente de correlação parcial (Veeriah, Durvasula, & Qi, 2015).

Time series analysis technique extrai com precisão os fenómenos de envelhecimento de determinadas séries temporais, analisa o comportamento do envelhecimento e classifica a sua gravidade (Ulanova, et al., 2015).

Phrase Segmentation consiste numa modelação estatística para inferir os limites de frases ocultas num texto. Para encontrar frases de forma eficiente, utiliza-se um modelo linguístico de frase *unigram*. Resumidamente, cada símbolo num modelo de linguagem *unigram* consiste numa ou mais palavras separadas por um espaço em branco. Ao codificar várias palavras dentro de um único *unigram*, o modelo de linguagem de frase é capaz de capturar relacionamentos de longa distância (Kıcıman & Richardson, 2015).

Canonicalization tem como objetivo perceber se duas pessoas estão a falar de um determinado assunto, mas utilizando diferentes expressões e palavras (Kıcıman & Richardson, 2015).

Multi Instance Learning centra-se em problemas de classificação onde rótulos são associados a conjuntos de instâncias, muitas vezes referidos como grupos, em vez de instâncias individuais. As etiquetas associadas aos grupos são assumidas como sendo rótulos de nível de instância não observados (Kotzias, Denil, De Freitas, & Smyth, 2015).

Naive Bayes Classifier refere-se a uma técnica probabilística capaz de prever o desempenho de um problema de classificação (Caballero Barajas & Akella, 2015).

GD-LDA Model tem com objetivo extrair tópicos de um conjunto de dados (Caballero Barajas & Akella, 2015).

Nonnegative Matrix Factorization consiste num conjunto de algoritmos de análise multivariada e álgebra linear onde uma matriz V é fatorizada em duas matrizes W e H, com a propriedade que as três matrizes não possuem elementos negativos. Esta não-negatividade torna as matrizes resultantes mais fáceis de inspecionar (Hayashi, Toyoda, & Kawarabayashi, 2015).

Conferência KDD2016 ano 2016 rank A*

Tabela 4 - Técnicas de análise de dados KDD2016

Ano 2016 rank A*

Referência	Artigo	Técnica/(as) de análise de dados	Tipo de dados
(Zhang, et al., 2016)	GMove: Group-Level Mobility Modeling using Geo-Tagged Social Media	Mobility Model	Não estruturado
(Huo, Nie, & Huang, 2016)	Robust and Effective Metric Learning Using Capped Trace Norm: Metric Learning via Capped Trace Norm	Metric Learning	Estruturado
(Chu, et al., 2016)	Finding Gangs in War from Signed Networks	Spectral Clustering	Estruturado/ Não estruturado
(Ye, Goebl, Plant, & Böhm, 2016)	FUSE: Full Spectral Clustering	Spectral Clustering	Estruturado/Não estruturado
(Chen & Joachims, 2016)	Predicting Matchups and Preferences in Context	Bradley–Terry Model	Estruturado
(Mu, et al., 2016)	User Identity Linkage by Latent User Space Modelling	SVM	Estruturado

Tabela 5 - Técnicas de análise de dados KDD2016 (continuação)

Ano 2016 rank A*

Referência	Artigo	Técnica/(as) de análise de dados	Tipo de dados
(Wei, Zheng, & Yang, 2016)	Transfer Knowledge between Cities	Multi-View Discriminant Transfer Learning	Estruturado
(Borisyuk, Kenthapadi, Stein, & Zhao, 2016)	CaSMoS: A Framework for Learning Candidate Selection Models over Structured Queries and Documents	Non-negative Coefficient Constrained Boundary Algorithm	Estruturado
(Li, Yao, Tang, Fan, & Tong, 2016)	QUINT: On Query-Specific Optimal Networks	Link Prediction	Semiestruturado
(Wang, Chen, Fei, Liu, & Emery, 2016)	Targeted Topic Modeling for Focused Analysis	Targeted Topic Model	Não estruturado

Mobility Model representa o movimento de um utilizador, a sua localização, velocidade entre outras características. Modelos deste tipo são frequentemente utilizados para fins de simulação, quando novas técnicas de comunicação ou navegação são investigadas. Os sistemas de gestão de mobilidade dos sistemas de comunicação móveis utilizam modelos de mobilidade para prever as futuras posições dos utilizadores (Zhang, et al., 2016).

Metric Learning tem como objetivo aprender a melhor projeção linear ou não linear para recursos de alta dimensão de restrições supervisionadas ou fracamente supervisionadas (Huo, Nie, & Huang, 2016).

Spectral Clustering consiste numa técnica que utiliza o *spectrum* de uma matriz de similaridade dos dados para realizar a redução da dimensionalidade antes do agrupamento em menor número de dimensões. A matriz de similaridade é fornecida como uma entrada e consiste em uma avaliação quantitativa da similaridade relativa de cada par de pontos no conjunto de dados (Ye, Goebl, Plant, & Böhm, 2016).

Bradley–Terry Model consiste num modelo probabilístico que prevê o resultado de uma comparação. Dado um par de indivíduos i e j retirados de uma população, ele estima a probabilidade de uma comparação par i > j seja verdadeira (Chen & Joachims, 2016).

Support Vector Machine consiste num conceito da ciência da computação para um conjunto de métodos de aprendizagem supervisionada que analisam os dados e reconhecem padrões, usado para classificação e análise de regressão (Mu, et al., 2016).

Multi-View Discriminant Transfer Learning tem como objetivo encontrar os vetores de pesos discriminantes ideais para cada *view,* de modo a que a correlação entre dados projetados em duas vistas seja maximizada, enquanto que tanto a discrepância de domínio e o desacordo de exibição são minimizados simultaneamente (Wei, Zheng, & Yang, 2016).

Non-negative Coefficient Constrained Boundary Algorithm consiste num algoritmo que realiza a modificação do algoritmo de descida de gradiente de modo a que os coeficientes negativos são ajustados para zero após cada passo de descida de gradiente. Este processo é repetido até que o procedimento de treinamento converge, ou o número de iteração atinge o limite (Borisyuk, Kenthapadi, Stein, & Zhao, 2016).

Link prediction pertence ao campo dos modelos de evolução da rede, que envolve o estudo de muitas redes sociais diferentes, tais como redes de citações, redes de comunicação, redes de conhecimento e redes de colaboração (Barabási & Albert, 1999).

Targeted topic model consiste num modelo de tópicos que, em vez ter a função de encontrar todos os tópicos de um documento, tem como objetivo encontrar um ou vários tópicos específicos para ajudar o utilizador a realizar análises mais profundas (Wang, Chen, Fei, Liu, & Emery, 2016).

ACM International Conference on Web Search and Data Mining (WSDM) rank A*

Tabela 6 - Conferência Web Search e Data Mining

Ano 2015 rank A*

Referência	Artigo	Técnica/(as) de análise de dados	Tipo de dados	Ano
(Wu & Ester, 2015)	FLAME: A Probabilistic Model Combining Aspect Based Opinion Mining and Collaborative Filtering	Text Mining	Não estruturado	2015
(Bi, et al., 2015)	Learning to Recommend Related Entities to Search Users	Probabilistic Three-Way Entity Model (TEM)	Semiestruturado	2015
(Zhuang & Young, 2015)	Leveraging In-Batch Annotation Bias for Crowdsourced Active Learning	Active Learning Algorithm	Não estruturado	2015
(Tang, Chang, Aggarwal, & Liu, 2015)	Negative Link Prediction in Social Media	Community Detection	Estruturado/ Não estruturado	2015

Tabela 7 - Conferência Web Search e Data Mining (continuação)

Ano 2015 rank A*

Referência	Artigo	Técnica/(as) de análise de dados	Tipo de dados	Ano
(Liu, Aggarwal, & Han, 2015)	On Integrating Network and Community Discovery	Community Detection Algorithm	Estruturado	2015
(Blanco, Ottaviano, & Meij, 2015)	Fast and Space-Efficient Entity Linking in Queries	Entity Linking	Semiestruturado	2015
(Chen, Hoi, Li, & Xiao, 2015)	SimApp: A Framework for Detecting Similar Mobile Applications by Online Kernel Learning	Kernels for Measuring App Similarity	Estruturado/ Não estruturado	2015
(Gao, Ma, & Chen, 2015)	Modeling and Predicting Retweeting Dynamics on Microblogging Platforms	Regression Models	Estruturado/ Não estruturado	2015
(Kokkodis, Papadimitriou, & Ipeirotis, 2015)	Hiring Behavior Models for Online Labor Markets	Bayesian Network	Estruturado/ Não estruturado/ Semiestruturado	2015
(Tran, Ceroni, Kanhabua, & Niederée, 2015)	Back to the Past: Supporting Interpretations of Forgotten Stories by Time-aware Re-Contextualization	Query Formulation	Estruturado	2015

Text Mining consiste numa variante de *Data Mining*, com o objetivo de descobrir novas informações previamente desconhecidas, extraindo automaticamente informações de diferentes recursos escritos. Um elemento-chave é a união das informações extraídas para formar novos fatos ou novas hipóteses a serem exploradas por meios mais convencionais de experimentação. Na pesquisa, o utilizador normalmente está a procurar algo que já é conhecido e foi escrito por outra pessoa. Com *Text Mining*, o objetivo é descobrir informações desconhecidas, algo que ninguém ainda sabe e por isso não poderia ter escrito (Senellart & Blondel, 2008); (Gupta & Lehal, 2009).

Probabilistic Three-way Entity Model(TEM) refere-se a um modelo probabilístico que fornece recomendações personalizadas de entidades relacionadas usando três fontes de dados: base de conhecimento, registo de cliques de pesquisa e *log* de painel de entidade. Especificamente, o *TEM* é capaz de extrair estruturas ocultas e capturar coorelações subjacentes entre usuários, entidades principais e entidades relacionadas. Além disso, o modelo *TEM* também pode explorar os sinais de clique derivados do *log* de painel de entidade (Bi, et al., 2015).

Active Learning Algorithm escolhe iterativamente itens de dados não rotulados para rotulagem por um oráculo externo e então dobram os novos dados rotulados de volta para o conjunto de dados de treinamento. Um modelo é treinado novamente no conjunto de treinamento atualizado. Quando é escolhido mais de um item de dados numa iteração, ele é chamado de aprendizado ativo em lote. Mais precisamente, supondo que o conjunto de dados não rotulado seja U, o conjunto rotulado seja

L e os rótulos de L estejam em yL. Um *Active Learning Algorithm* escolhe um subconjunto A ⊂ U com um dado com tamanho k e obtém seus rótulos yA de um oráculo. Os dados selecionados são então movidos de U e adicionados a L com seus rótulos yA concatenados a yL (Zhuang & Young, 2015).

Community Detection tem como objetivo procurar descrever a estrutura em larga escada de uma rede, dividindo os nós em comunidades, também chamados de blocos ou grupos, com base apenas no padrão de *links*. Essa tarefa é semelhante à de dados vetoriais de *cluster*, pois ambos procuram identificar grupos importantes dentro de um conjunto de dados (Peel, Larremore, & Clauset, 2016).

Entity Linking tem como objetivo a determinação da identidade das entidades mencionadas num determinado texto. Por exemplo, dada a frase "Paris é a capital de França", a ideia é determinar que "Paris" se refere à cidade de Paris e não à Paris Hilton ou qualquer outra entidade que poderia ser referida como "Paris". *Entity Linking* requer uma base de conhecimento contendo as entidades às quais as menções de entidade podem ser ligadas (Blanco, Ottaviano, & Meij, 2015).

Kernels em *machine learning,* é essencialmente uma função de mapeamento que transforma um "*low-dimensional space*" em um "*higher-dimensional space*". Uma função do *kernel* pode ser pensada como uma função de similaridade *parwise* (Chen, Hoi, Li, & Xiao, 2015).

Regression Models consiste em modelos estatísticos para estimar as relações entre variáveis. Inclui muitas técnicas para modelar e analisar várias variáveis, quando o foco está na relação entre uma variável dependente e uma ou mais variáveis independentes. A análise de regressão é amplamente utilizada para previsão (Gao, Ma, & Chen, 2015).

Querys Formulations consiste em gerar um conjunto de *queries SQL* Qd para um dado documento d para recuperar candidatos de contextualização como *input* para *re-ranking*. Há dois métodos de *query formulations,* um utilizando o documento para ser contextualizado como um "gerador" de consultas e outro utilizando ganchos de contextualização como "gerador" (Tran, Ceroni, Kanhabua, & Niederée, 2015).

Conferência International Journal of Big Data Intelligence (IJBDI)

Tabela 8 - Conferência International Journal of Big Data Intelligence

Ano 2014 a 2017

Referência	Artigo	Técnica/(as) de análise de dados	Tipo de dados	Ano
(Zalmout & Ghanem, 2016)	Multivariate adaptive community detection in Twitter	Bayesian Classification	Estruturado/Não estruturado/ Semiestrutrado	2016
(Oliveira, Barbar, & Soares, 2016)	Computer network traffic prediction: a comparison between traditional and deep learning neural network	Multilayer Perceptron, (MLP)	Estruturado/ Semiestruturado/Não estruturado	2016
(Wlodarczyk & Hacker, 2014)	Current trends in predictive analytics of big data	Predictive Analytics	Estruturado	2014
(Lomotey & Deters, 2015)	Unstructured data mining: use case for CouchDB	Inference-Based Apriori with a Bayesian Component, The Hidden Markov Model, Bernoulli Process	Não estruturado	2015

Tabela 9 - Conferência International Journal of Big Data Intelligence (continuação)

Ano 2014 a 2017

Referência	Artigo	Técnica/(as) de análise de dados	Tipo de dados	Ano
(Sharma, et al., 2015)	Classification and comparison of NoSQL big data models	Classification of NoSql Data Models	Não estruturado	2015
(Geerdink, 2015)	A reference architecture for big data solutions - introducing a model to perform predictive analytics using big data technology	Data Discovery	Não estruturado	2015
(Alshareef, Bakar, Hamdan, Abdullah, & Alweshah, 2015)	A case-based reasoning approach for pattern detection in Malaysia rainfall data	Association Rule Learning	Estruturado/ Não estruturado	2015
(Makrynioti, et al., 2017)	PaloPro: a platform for knowledge extraction from big social data and the news	Text mining: Entity Recognition,	Semiestruturado/ Não estruturado	2017
(Yang, Milosevic, & Cao, 2017)	Optimising column family for OLAP queries in HBase	OLAP Queries	estruturado	2017
(Dimitrakopoulos, Chatzigiannakis, & Tsitouras, 2017)	A knowledge-based integrated framework for increasing social management intelligence	Sentiment Analysis	Semiestruturado/ Não estruturado	2017
(Makrynioti, et al., 2017)	PaloPro: a platform for knowledge extraction from big social data and the news	Text mining: Entity Recognition,	Semiestruturado/ Não estruturado	2017

Multilayer perceptron(MLP) consiste num modelo de rede neuronal artificial que mapeia um conjunto de dados de *input* para um conjunto de *output* definido. *MLP* consiste em várias camadas de nós num gráfico direcionado, com cada camada totalmente conectada ao próximo. Exceto para os nós de entrada, cada nó é um neurónio com uma função de ativação não-linear. *MLP* utiliza uma técnica de aprendizagem supervisionada chamada de *backpropagation* para treinar a rede (Oliveira, Barbar, & Soares, 2016).

Predictive analytics engloba uma variedade de técnicas estatísticas de modelação preditiva, *machine learning* e *data mining* que analisam fatos atuais e históricos para realizar previsões sobre eventos futuros ou de outros fatores desconhecidos (Nyce & CPCU, 2007).

Classification of NoSql Data models consiste na classificação de três modelos distintos com características diferentes (Sharma, et al., 2015):

- *Document-Oriented Store:* nesta categoria, os dados formatados ou codificados em *XML, YAML (do inglês YAML Ain't Markup Language), JSON (do inglês JavaScript Object Notation), ou BSON (*do inglês *Binnary JSON)* são encapsulados em documentos e a organização desses documentos é dependente da implementação. No armazenamento, a cada documento é atribuído uma chave exclusiva que serve como identificador exclusivo para esse documento específico e os modelos de dados estão equipados com *APIs* ou algum sistema de processamento de consultas para aceder os documentos. Alguns exemplos de modelos de dados desta categoria incluem o MongoDB, CouchDB, OrientDb, Couchbase, MarkLogic, RavenDB, Clouddant, GemFire, ente outros.
- *Graph Data Model:* estes modelos de dados armazenam os dados ao longo de nós, bordas e propriedades de uma estrutura gráfica e proporcionam adjacência independente de índice, onde cada elemento é conectado aos seus elementos adjacentes através de ponteiros diretos. Isso evita qualquer pesquisa para indexação. Alguns *Graph models* são o Neo4J, OrientDB, InfiniteGraph, Allegro, Virtuoso e Stardog.
- *Key-value Store:* ao mais baixo nível, o armazenamento de *key-value* pode ser construído por um *array* associativo – mapa ou dicionário. Esse modelo de dados, armazena os dados como uma coleção de pares *key-values,* onde a chave serve como a chave primária que não pode ser duplicada na mesma coleção. O modelo de dados *key-value*

consiste num modelo de dados não triviais mais básicos que sustentam o desenvolvimento de modelos de dados mais complexos.

Association Rule Learning é utilizado para descobrir elementos que ocorrem em comum dentro de um determinado conjunto de dados. *Association rules* tem como premissa encontrar elementos que implicam a presença de outros elementos numa mesma transação, ou seja, encontrar relacionamentos ou padrões frequentes num dado conjunto de dados. Existem diversos algoritmos que realizam buscas de *Association Rules* em base de dados, como por exemplo: *Apriori, Partition, Eclat e FP-Growth* (Alshareef, Bakar, Hamdan, Abdullah, & Alweshah, 2015).

Entity recognition consiste numa sub-tarefa de extração de informações com o objetivo de localizar e classificar determinadas entidades em categorias pré-definidas, como por exemplo nomes das pessoas, organizações locais, expressões de tempos, quantidades, entre outros (Makrynioti, et al., 2017).

OLAP Queries refere-se a uma abordagem para responder a consultas analíticas de modelos multidimensionais na área de computação. *OLAP* permite às organizações um método de acesso, visualização e análise de dados corporativos com alta flexibilidade e desempenho. Como exemplo de aplicações de *OLAP*, temos relatórios de negócios para vendas, de marketing, relatórios de gestão, análise de processos de negócio, orçamento e previsão, relatórios financeiros, entre outros (Kimball & Ross, 2011).

Sentiment analysis também conhecido como *Opinion mining,* consiste na análise de opiniões, sentimentos, avaliações, atitudes e emoções das pessoas para uma determinada entidade como produtos, serviços, organizações, indivíduos, problemas, tópicos e os seus atributos. As opiniões são um ponto central em quase todas as atividades humanas porque são os principais influenciadores dos nossos comportamentos, e devido a esse facto tem existido maior investigação na área *de Natural Language Processing, data mining, web mining* e extração de informação desde o ano 2000 (Liu B. , 2012).

De forma a sintetizar as técnicas ilustradas acima, foram elaboradas as tabelas 10, 11, 12,13 e 14 em que para cada formato de dados, estruturado, semiestruturado e não estruturado, são apresentadas as técnicas que podem ser utilizadas para efetuar análise de dados.

Tabela 10 - Técnicas para dados estruturados

Referência	Técnicas
(Zhao, Liu, & Cox, 2014)	Sparse Support Vector Machine
(Anagnostopoulos & Triantafillou, 2014)	Sequential Multivariate Regression Imputation
(Veeriah, Durvasula, & Qi, 2015)	Time-Series Shapelets
	Cross-Correlation
	Granger Causality
	Information Gain
	Trained Classifier
	Random Score
	Time-Series Shapelets
	Cross-Correlation
	Generalized Transfer Entropy (GTE)
	Partial Correlation Statistics
(Ulanova, et al., 2015)	Time Series Analysis Technique
(Kotzias, Denil, De Freitas, & Smyth, 2015)	Multi Instance Learning (MIL)
(Caballero Barajas & Akella, 2015), (Kokkodis, Papadimitriou, & Ipeirotis, 2015), (Zalmout & Ghanem, 2016)	Naive Bayes Classifier
(Hayashi, Toyoda, & Kawarabayashi, 2015)	Latent Dirichlet Allocation
(Hayashi, Toyoda, & Kawarabayashi, 2015)	Nonnegative Matrix Factorization
(Huo, Nie, & Huang, 2016)	Metric Learning
(Chu, et al., 2016), (Ye, Goebl, Plant, & Böhm, 2016)	Spectral Clustering
(Chen & Joachims, 2016)	Bradley–Terry Model
(Mu, et al., 2016)	SVM
(Borisyuk, Kenthapadi, Stein, & Zhao, 2016)	Multi-View Discriminant Transfer Learning
(Liu, Aggarwal, & Han, 2015)	Community Detection Algorithm
(Chen, Hoi, Li, & Xiao, 2015)	Kernels for Measuring App Similarity
(Tran, Ceroni, Kanhabua, & Niederée, 2015)	Query Formulation
(Gao, Ma, & Chen, 2015)	Regression Models
(Oliveira, Barbar, & Soares, 2016)	Multilayer Perceptron (MLP)

Tabela 11 - Técnicas para dados estruturados (continuação)

Referência	Técnicas
(Wlodarczyk & Hacker, 2014)	Predictive Analytics
(Alshareef, Bakar, Hamdan, Abdullah, & Alweshah, 2015)	Association Rule Learning
(Yang, Milosevic, & Cao, 2017)	OLAP Queries

Tabela 12 - Técnicas para dados semiestruturados

Referência	Técnicas
(Dong, et al., 2014)	HTML Trees (DOM)
(Li, Yao, Tang, Fan, & Tong, 2016), (Zhou, Liu, & Buttler, 2015)	Link Prediction
(Bi, et al., 2015)	Probabilistic Three-way Entity Model (TEM)
(Blanco, Ottaviano, & Meij, 2015)	Entity Linking
(Caballero Barajas & Akella, 2015), (Kokkodis, Papadimitriou, & Ipeirotis, 2015), (Zalmout & Ghanem, 2016)	Naive Bayes Classifier
(Oliveira, Barbar, & Soares, 2016)	Multilayer Perceptron, (MLP)
(Makrynioti, et al., 2017)	Entity Recognition
(Makrynioti, et al., 2017), (Dimitrakopoulos, Chatzigiannakis, & Tsitouras, 2017)	Sentiment Analysis

Tabela 13 - Técnicas para dados Não estruturados

Referência	Técnicas
(Sudhof, Goméz, Maas, & Potts, 2014)	Conditional Random Fields (CRFs) as a Modeling Technique
(Dong, et al., 2014), (Mukherjee, Weikum, & Danescu-Niculescu-Mizil, 2014), (Shashidhar, Pandey, & Aggarwal, 2015)	Natural Language Processing
(Chen & Liu, 2014), (Wang, Chen, Fei, Liu, & Emery, 2016)	Topic Modeling
(Kurashima, Iwata, Takaya, & Sawada, 2014)	Probabilistic Latent Semantic Visualization
(Schubert, Weiler, & Kriegel, 2014)	First Story Detection
(Zhao, Liu, & Cox, 2014)	Sparse Support Vector Machine
(Nagarajan, et al., 2015)	Event Extraction
(Kıcıman & Richardson, 2015)	Phrase Segmentation
(Kıcıman & Richardson, 2015)	Canonicalization
(Caballero Barajas & Akella, 2015), (Kokkodis, Papadimitriou, & Ipeirotis, 2015), (Zalmout & Ghanem, 2016)	Naive Bayes Classifier
(Hayashi, Toyoda, & Kawarabayashi, 2015)	Latent Dirichlet Allocation

Tabela 14 - Técnicas para dados Não estruturados (continuação)

Referência	Técnicas
(Hayashi, Toyoda, & Kawarabayashi, 2015)	Nonnegative Matrix Factorization
(Zhang, et al., 2016)	Mobility Model
(Chu, et al., 2016), (Ye, Goebl, Plant, & Böhm, 2016)	Spectral Clustering
(Wu & Ester, 2015)	Text Mining
(Zhuang & Young, 2015)	Active Learning Algorithm
(Tang, Chang, Aggarwal, & Liu, 2015), (Liu, Aggarwal, & Han, 2015)	Community Detection
(Chen, Hoi, Li, & Xiao, 2015)	Kernels for Measuring App Similarity
(Gao, Ma, & Chen, 2015)	Regression Models
(Zalmout & Ghanem, 2016)	Multilayer Perceptron (MLP)
(Geerdink, 2015)	Data Discovery
(Alshareef, Bakar, Hamdan, Abdullah, & Alweshah, 2015)	Association Rule Learning
(Makrynioti, et al., 2017)	Entity Recognition
(Makrynioti, et al., 2017), (Dimitrakopoulos, Chatzigiannakis, & Tsitouras, 2017)	Sentiment Analysis

2.3 Ferramentas de Análise de Dados

De seguida, é apresentado um conjunto de cinco ferramentas de análise de dados, com o objetivo de perceber que tipo de funcionalidades estas mesmas oferecem.

Watson Analytics[1]

Desenvolvido pela *IBM, Watson Analytics* consiste num serviço inteligente de análise e visualização de dados que se pode utilizar para descobrir rapidamente padrões e significados nos seus dados. Com a descoberta guiada de dados, a análise preditiva automatizada e as capacidades cognitivas, como o diálogo em linguagem natural, possuem a capacidade de interagir com os dados conversacionalmente para obter as respostas que se pretende.

Funcionalidades:

- *Smart data discovery*: possui a capacidade de encontrar padrões relevantes num conjunto de dados, com base em palavras referidas pelo utilizador, através da técnica de processamento da linguagem natural;
- *Simplified analysis:* permite perceber rapidamente um conjunto de informações de um conjunto de dados utilizando automação*;*
- *Accessible advanced analytics:* permite conectar-se a um conjunto de dados sem assumir o refinamento ou preparação de dados complexos. Esta ferramenta permite eliminar a

[1] https://www.ibm.com/br-pt/marketplace/watson-analytics

complexidade e tarefas demoradas, enquanto o utilizador se apoia as suas decisões com base em dados que pode confiar*;*

- ***Self-service dashboards***: esta ferramenta permite compartilhar um conjunto de ideias num *dashboard* que pode ser construído facilmente a partir de visualizações que o utilizador salve durante a descoberta de dados*;*

Oracle[2]:

Oracle Analytics Cloud permite fornecer análise de negócios para dados tradicionais e para *Big Data* em toda a organização. Com base em tecnologias comprovadas da *Oracle* e infraestrutura *Cloud, Oracle Analytics Cloud* consiste num portefólio abrangente de ofertas para a *Cloud*, abrangendo todas as necessidades dos consumidores em *Business intelligence, Big Data Analytics,* e *SaaS* (do inglês *Software as a Service)*.

SCM Analytics[3]

Para executivos:

- ***Proactive problem solving****:* permite passar de uma tomada de decisão reativa a proativa, através da visibilidade precoce na realização de pedidos, inventário, entre outros*;*
- ***Real Time Analysis***: possui a capacidade de obter informações em tempo real sobre balanço de inventário, poupança feita através de políticas de desconto, entre outros de forma a maximizar a satisfação do cliente e a receita da organização;
- ***Innovation Management****:* permite acompanhar e monitorizar o desempenho da gestão de inovação da organização, como por exemplo acompanhar a transformação da execução de ideias e novos produtos em ofertas lucrativas para clientes*;*
- ***Supply Chain Otimization****:* possui a capacidade de obter clareza sobre o estado de pedidos abertos, potenciais coleções, faturas em espera, entre outros, de forma a efetivamente melhorar a satisfação do cliente;

[2] https://cloud.oracle.com/en_US/saas

[3] https://cloud.oracle.com/en_US/scm-analytics/features

Para utilizadores de nível operacional:

- ***Backlog:*** permite obter visibilidade clara sobre o número de linhas em execução no momento em atraso*;*
- ***Insight-Driven Supply Chain Management:*** possui a capacidade de realizar uma decisão informada através da visibilidade em tempo real dos saldos para itens, pedidos e ações pendentes;
- ***Late and On-hold Orchestration:*** permite obter informações sobre o número de pedidos em atraso e suspensos no sistema;
- ***Product Governance:*** permite obter informações sobre atividades de gestão de produtos, como importações de lotes de itens, pedidos de alteração de estado, entre outros;

Microsoft[4]:

Desenvolvido pela *Microsoft, Power BI* consiste num conjunto de ferramentas de análise de dados, para poder analisar o estado de uma organização e compartilhar ideias. O *Power BI* pode unificar todos os dados de uma organização, quer seja na nuvem ou localmente. Usando os *gateways* do *Power BI*, você pode conectar as bases de dados *SQL Server*, modelos do *Analysis Services* e muitas outras fontes de dados aos mesmos painéis no *Power BI*.

Power BI fornece os seguintes serviços:

- ***Dashboards****:* permite criar um conjunto de visualizações sobre dados;
- ***New visualizations:*** possui a capacidade de criar um conjunto de visualizações diferentes de formas a ilustrar uma ideia sobre diferentes perspetivas*;*
- ***Connectors for popular SaaS Services:*** possui a capacidade de interligar a ferramenta *PowerBI* com um conjunto de aplicações *SaaS* populares, por exemplo *Salesforce*, *Zendesk*, *Marketo*, *SendGrid*, *GitHub*, entre outros*;*
- ***Mobile app for iPad:*** permite utilizar o serviço de *PowerBI, PowerBI Report Server* e *Reporting Services* no iPad*;*
- ***Live connectivity to SQL Server Analysis Services:*** permite criar uma ligação com o *SQL Server Analysis Services;*

[4] https://powerbi.microsoft.com/pt-br/

- ***Power BI Designer**:* esta aplicação combina o *Power Query, Power Pivot Data Model* e *Power View* em uma experiência que permitirá que os utilizados criem os seus elementos do *PowerBI* de forma offline e, em seguida, carregar quando pretenderem no serviço *PowerBI;*

TIBCO Spotfire[5]:

Desenvolvido por *TIBCO Spotfire*, *TIBCO Spotfire Cloud* consiste num software de análise de dados como um serviço para prospeção de dados.

TIBCO Spotfire Cloud fornece os seguintes serviços de análise:

- ***Smart Visual Data Discovery:*** permite criar uma análise visual inteligente, utilizando um mecanismo de recomendação orientado por *AI* (do inglês *Artificial Intelligence*), e um conjunto de ferramentas de descoberta de dados;
- ***Immersive Data Wrangling:*** os recursos de preparação de dados ajudam a moldar, enriquecer e transformar os dados, assim como identificar sinais para *dashboards*. *Data Wrangling* permite ao utilizador detetar rapidamente dados fora do padrão, inconsistentes e deficiências;
- ***Preditictive Analytics:*** permite criar modelos preditivos com recurso ao uso da ferramenta R e conectores para *PMML*, *H2O* e *SparkML, m*as não só, também com a capacidade de se conectar a *RServe*, *SAS* e *Mathlab*, permite ao utilizador criar visualizações à sua medida;
- ***Location Analytics:*** permite clarificar instantaneamente a localização das análises feitas pelo utilizador. *Spotfire* possui capacidades de mapeamento de multicamadas capazes de se conectar a muitos serviçõe de mapas, para que o utilizador possa escolher os mapas que pretende para ilustrar uma ideia;

[5] https://spotfire.tibco.com/overview

3 Descrição do Trabalho Realizado

Neste ponto, irá ser ilustrado o artefacto construído para resolver o problema proposto nesta dissertação. Neste artefacto, foram utilizadas quatro técnicas de análise de dados, referidas no ponto 3.1, foram utilizados dois *datasets* explicados no ponto 3.2. As ferramentas utilizadas são apresentadas no ponto 3.3. No ponto 3.4 é apresentado a arquitetura global do artefacto. A forma como foi realizada a extração dos dados não estruturados e semiestruturados é apresentada no ponto 3.5. Para avaliar a performance do artefacto criado, no ponto 3.6 é ilustrado um teste de forma a provar que é possível retirar informações de dados não estruturados e semiestruturados. No ponto 3.7, é ilustrado a plataforma analítica criada para se poder proceder a análise de dados, utilizando cubos *OLAP*.

3.1 Técnicas de análise utilizadas

Neste artefacto, foram utilizadas quatro técnicas de análise sendo elas:

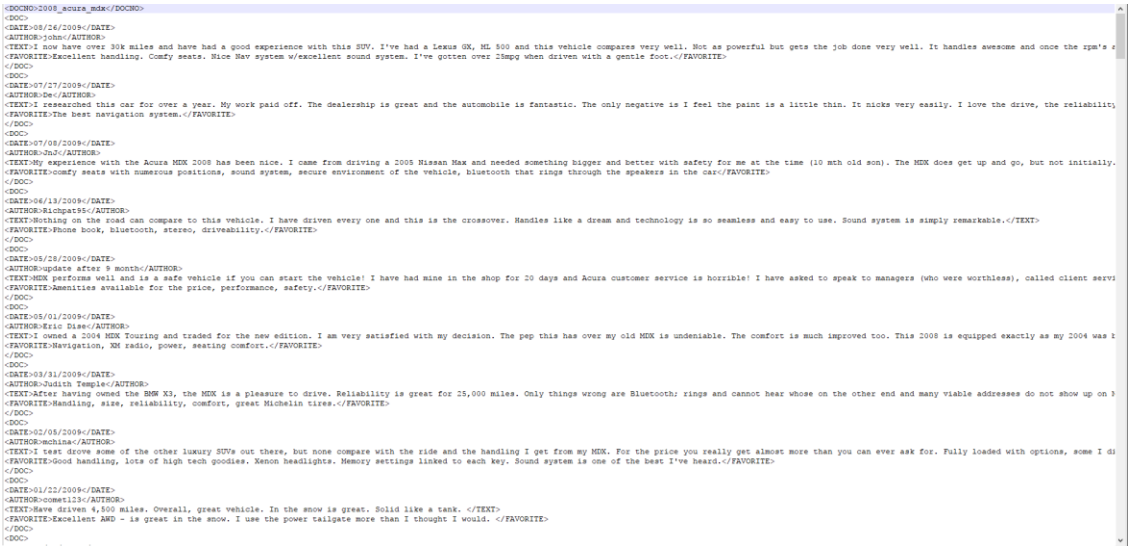
- **Processamento da Linguagem Natural**: utilizada para encontrar um conjunto de padrões, nomeadamente um conjunto de relações, palavras-chave e entidades;
- **Análise de Sentimento**: com o objetivo de definir a análise de sentimento associada a um determinado padrão;
- **Análise de Emoção:** com o objetivo de definir a análise de emoção associada a um determinado padrão;
- **Reconhecimento de imagens:** com o propósito de reconhecer um conjunto de imagens e associar a uma categoria;

3.2 Datasets

O *dataset* usado neste artefacto foi utilizado pelos autores Ganesan & Zhai (2012) num estudo sobre classificação de entidades baseadas em opiniões e é composto por 3 anos de comentários sobre carros, isto é, 2007,2008 e 2009. Este *dataset* possui até 250 carros de diferentes modelos por ano. O conteúdo textual está dividido entre o campo de favoritos, nomes dos autores,

datas, e a revisão textual propriamente dita sobre o carro. Sendo este *dataset* considerado semiestruturado.

O formato do ficheiro que possui os comentários é ilustrado na figura 2.



Retirado da ferramenta *Postman*

Figura 2 - Formato Dataset Comentários

Nesta experiência apenas foram utilizados os comentários alusivos aos carros do ano 2008.

Em complemento do *dataset* de comentários de carros referenciado atrás, foi ainda utilizado um *dataset* composto por imagens de carros, que representa modelos de carros do ano 2008, para se proceder a classificação da categoria de carro de cada modelo. Sendo este *dataset* considerado como dados não estruturados.

3.3 Ferramentas Utilizadas

A escolha das ferramentas prendeu-se essencialmente devido às parcerias que a Universidade do Minho possui com as empresas do mercado. Nomeadamente, a parceria com a *Microsoft,* foi possível o acesso as ferramentas *SQL Server*, *SSAS* e *SSIS,* possuindo uma licença sem limite de tempo. Também foi utilizado a parceria com a *IBM*, para ter acesso as ferramentas *Natural Language Understanding*, *Watson Knowledge Studio* e *Visual Recognision*, possuindo uma licença válida por 6 meses.

De resto foi utilizado a ferramenta *Talend* com uma licença de 30 dias, e a ferramenta *Postman* com uma licença grátis sem limite de tempo.

Natural Language Understanding IBM

A ferramenta *Natural Languagem Understanding* da *IBM*, mais concretamente da *IBM Bluemix Watson*, permite a análise de texto através do processamento de linguagem natural.

Este serviço permite analisar texto não estruturado, extrair meta dados como: entidades, palavras-chaves, relações, o sentimento, a emoção, entre outros. Através de um modelo de *machine-learning* personalizado de anotações criados na ferramenta *Watson Knowledge Studio*, é possível personalizar o serviço para identificar entidades e relações específicas para o caso de estudo que se pretende estudar.

Esta ferramenta foi utilizada na experiência para retirar um conjunto de informações: entidades, palavras chaves, relações, análise de sentimento e emoção; dos comentários alusivos a um determinado modelo de carro.

Watson Knowledge Studio IBM

Watson Knowledge Studio consiste numa ferramenta capaz de anotar texto não estruturado, e utiliza essas anotações para criar um modelo de *machine-learning* customizado capaz de perceber a linguagem do domínio que se pretende estudar. O resultando deste modelo consiste num algoritmo capaz de aprender padrões e reconhecer esses padrões numa grande coleção de documentos.

Na experiência, esta ferramenta foi utilizada como suplemento da ferramenta *Natural Language Understanding*, e tinha como objetivo reconhecer um conjunto de padrões através de um modelo *machine learning*.

Visual Recognision IBM

A ferramenta *Visual Recognision* é fornecida pela *IBM*, mais concretamente pela *IBM Bluemix Watson*, e tem como objetivo reconhecer o conteúdo das imagens. Através de um classificador customizado é possível categorizar uma coleção de imagens semelhantes.

Nesta experiência, a ferramenta *Visual Recognision* foi utilizada para associar uma imagem a uma categoria de carros, e tal foi possível utilizando um classificador customizado para as categorias que se pretendiam reconhecer.

Postman

Esta aplicação tem como principal objetivo fornecer uma interface gráfica capaz de comunicar, através de pedidos do tipo *POST* ou *GET*, com outras aplicações através de *APIs*. Possui uma interface amigável, e é fácil de utilizar.

A ferramenta *Postman* foi utilizada para realizar os pedidos do tipo *POST* para a ferramenta *Visual Recognision* (para classificar um modelo de carro a uma categoria), e para a ferramenta *Natural Language Understanding* (para obter informações sobre os comentários associados a um carro). As informações obtidas foram guardadas no formato *JSON*.

SQL Server

A versão utilizada *SQL Server Enterprise Edition 2014* e serviu como repositório de dados, onde foi guardado a base de dados estágio e os Cubos *OLAP*.

SSIS

O *SQL Server Integration Services* (*SSIS*) foi utilizado para desenvolver funcionalidades de *ETL*.

SSAS

O *SQL Server Analysis Services* (SSAS) foi utilizado para criar os cubos *OLAP*.

Talend

A ferramenta *Talend* consiste num produto de integração de dados *Open Source* desenhado para combinar, converter e transformar dados para as necessidades dos utilizadores.

Esta ferramenta foi utilizada, visto que o *SSIS* não possuía a funcionalidade de lidar com dados no formato *JSON*. Portanto, foi utilizado esta ferramenta para tarefas de *ETL*, nomeadamente transformar dados do formato *JSON* para o formato tabular, e inserir numa base de dados.

3.4 Arquitetura Global do Artefacto

Neste ponto, irá ser mostrado a arquitetura global do artefacto realizada na figura 3.

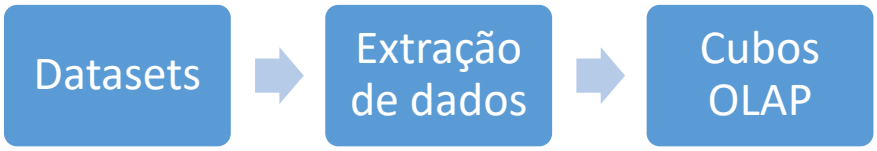


Figura 3 - Arquitetura Global do Artefacto

Como ilustrado na figura 3, o artefacto é composto por 3 componentes distintos. O primeiro consiste nos *datasets* de entrada referido no ponto 3.2. O ponto "extração de dados", explicado no ponto 3.5, passa por transformar os dados recebidos dos *datasets*, comentários sobre carros, e de imagens a ilustrar o modelo dos carros, em dados estruturados. De seguida os dados vão ser inseridos numa base de dados, e posteriormente procedera-se a criação de cubos *OLAP*, explicado no ponto 3.7.

3.5 Extração de Dados

De seguida, vai ser ilustrado o ponto de extração de dados, e qual o papel de cada um dos componentes ilustrados na figura 4.

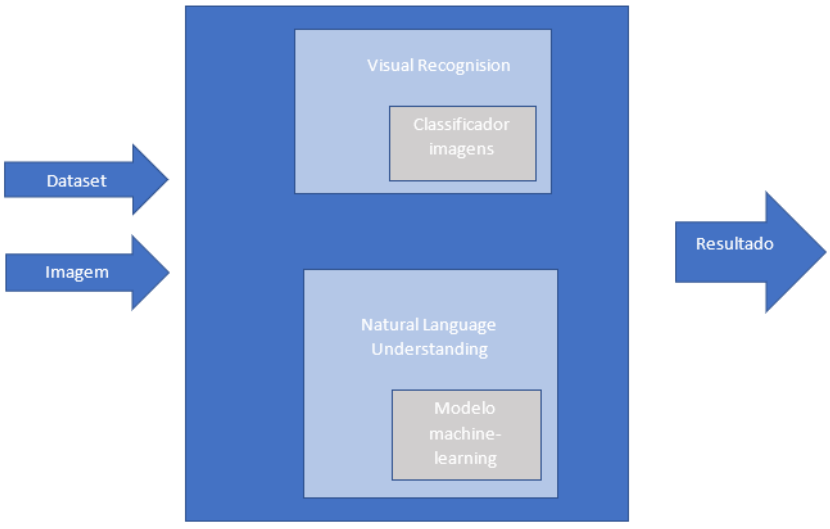


Figura 4 - Arquitetura Extração de dados

Na figura 4, é apresentado como dados de entrada um *dataset* contendo comentários sobre um modelo de carro, assim como uma imagem a ilustrar o modelo de carro respetivamente.

A imagem obtida como dado de entrada vai ser classificada recorrendo a ferramenta *Visual Recognision*. Nesta ferramenta, foi necessário criar um classificador personalizado de forma a poder reconhecer a que categoria pertence o modelo do carro.

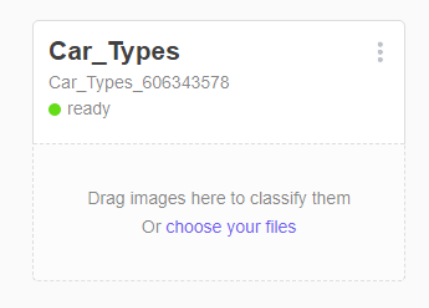


Figura 5 - Classificador Visual Recognision

Retirado da ferramenta Visual Recognision

Para poder criar o classificador ilustrado na figura 5, foi necessário criar um conjunto de categorias, apresentadas alguns exemplos na seguinte tabela. Em anexo é apresentado a tabela completa.

Tabela 15 - Categorias de Carros

Categorias
Crossover_SUV
Full_Size_Sedan
Mid_Size_Sedan
Compact_executive_luxury
Compact_Family

Para cada categoria, foi adicionado um conjunto de 10 imagens distintas de forma a ferramenta poder treinar o classificador. Depois do classificador estar criado, é fornecido uma chave para poder ser utilizada com o recurso a uma *API*. O formato da chave foi o seguinte: "Car_Types_606343578".

Por outro lado, de forma a poder retirar as relações, entidades, palavras-chave, sentimento e a emoção do *dataset* que possui comentários sobre carros, foi utilizado a ferramenta *Natural Language Understanding*.

Para a ferramenta *Natural Language Understanding* conseguir encontrar o tipo de informações que se pretendia do *dataset*, comentários sobre carros, foi necessário recorrer a ferramenta *Watson Knowledge Studio*.

Na ferramenta *Watson Knowledge Studio*, foi possível criar um modelo de *machine-learning* capaz de anotar um conjunto de padrões relevantes para esta experiência.

O *workflow* para a criação do modelo é o seguinte ilustrado na figura 6.

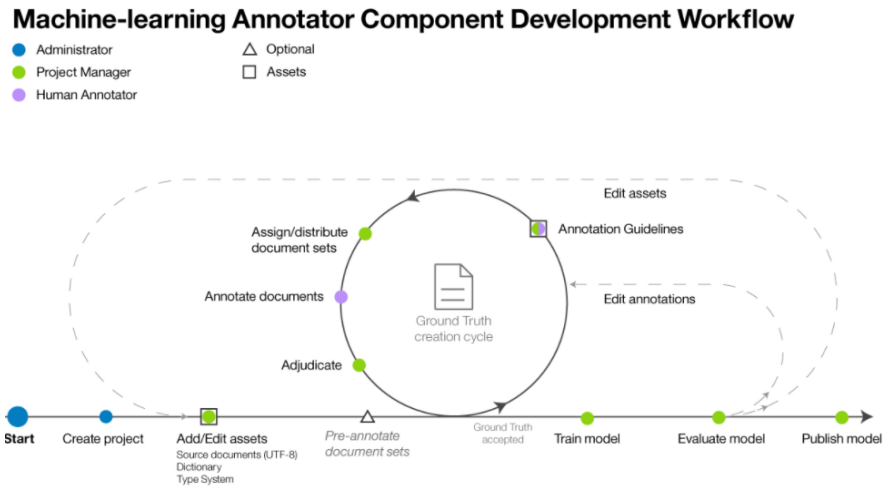


Figura 6 - Passos para Criar Modelo Machine Learning

Retirado da ferramenta *Watson Knowledge Studio*

O primeiro passo para a criação do modelo *machine-learning*, passa pela criação do projeto, e de seguida estabelecer os tipos de entidades que se pretende identificar, ilustrado na figura 6 como *Type System*. Apresenta-se alguns exemplos de entidades utilizadas na tabela 16. Em anexo, é apresentado a lista completa.

Tabela 16 - Entidades criadas

Entidades
Suspension
Oil_change
Rims
Weels
Sport
Luxury

Depois das entidades ficarem determinadas, foram definidas as relações que se pretendiam identificar. Nesta experiência as relações criadas foram as de "*Favorite*" e "*Problem*" e foram associadas às entidades.

O próximo passo, passou por adicionar um conjunto de 8 documentos dos carros do ano 2008, i.e., ficheiros que constituem o *dataset*, com o objetivo de anotar as respetivas entidades e relações. Devido a limitações da ferramenta, só foi possível utilizar excertos dos 8 documentos.

Para poder anotar os documentos foram criadas 2 tarefas distintas ilustradas na figura 7, contendo cada uma, 4 documentos distintos.

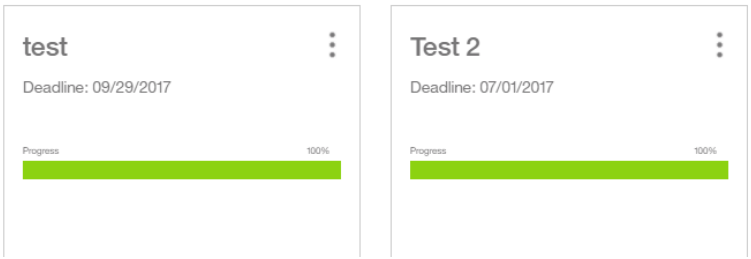


Figura 7 - Tabelas de anotação

Retirado da ferramenta *Watson Knowledge Studio*

A interface que a ferramenta *Watson Knowledge Studio* apresenta para anotação das entidades está ilustrada na figura 8.

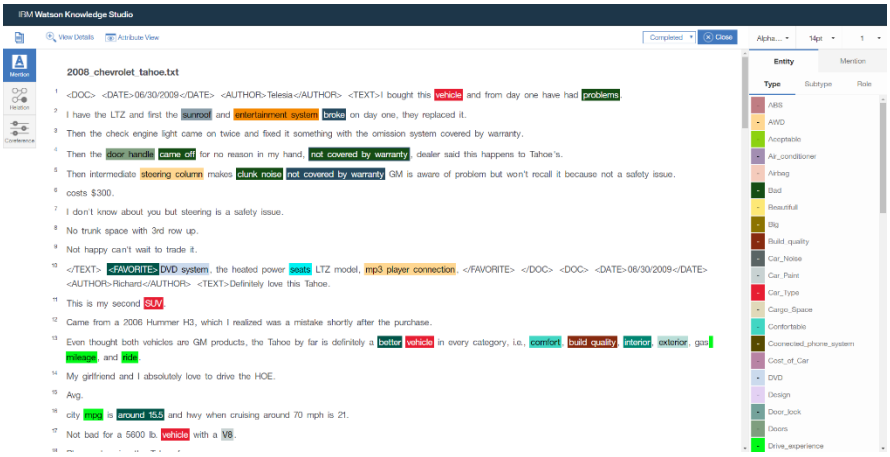


Figura 8 - Interface anotar entidades

Retirado da ferramenta *Watson Knowledge Studio*

No lado direito da figura 8, temos as entidades que podemos utilizar, e ao clicar no texto ele vai identificar essa palavra ou conjunto de palavas como referindo a essa entidade.

A interface para anotação das relações está ilustrada na figura 9.

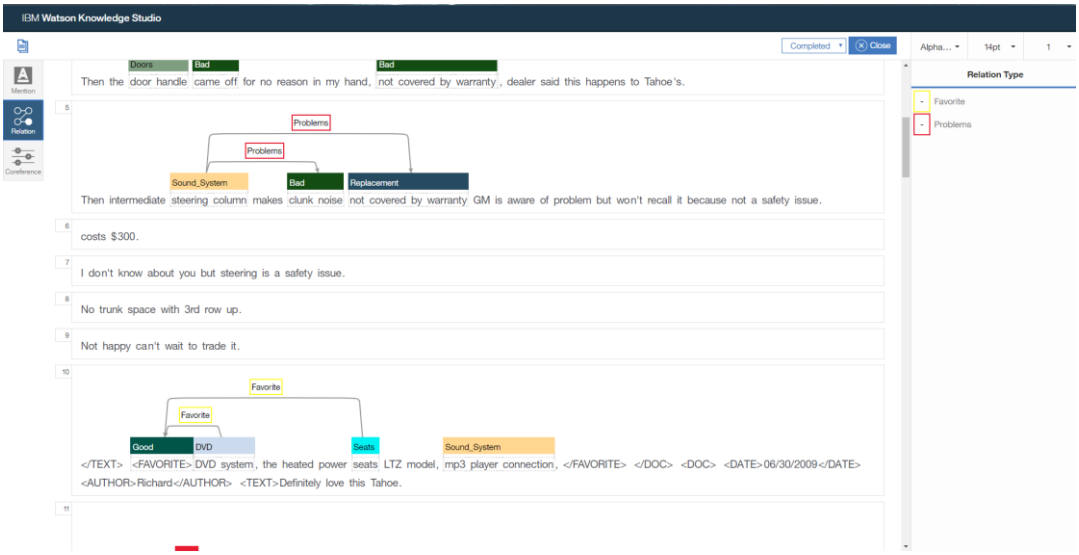


Figura 9 - Interface anotar relações

Retirado da ferramenta *Watson Knowledge Studio*

No lado direito da figura 9, estão ilustradas as diferentes relações criadas, e no texto ao selecionar duas entidades é possível estabelecer uma relação, neste caso específico atribuição de uma entidade como favorito, ou atribuição de uma entidade como um problema.

As anotações quer das entidades, como das relações dos documentos identificados nas duas tarefas ilustradas na figura 7, vão ser classificadas pelo modelo como a criação do *Ground of Truth* referenciado na figura 6.

De seguida, o modelo criado é treinado e avaliado, e por fim publicado. A publicação do modelo cria uma chave denominada "*model_id*" ilustrada na figura 10, que vai ser associada à ferramenta *Natural Language Understanding*.

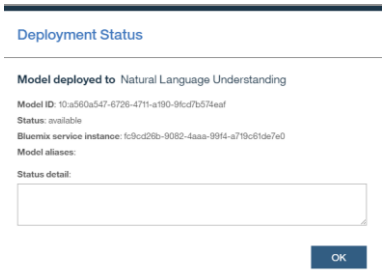


Figura 10 - Chave do modelo Machine Learning

Retirado da ferramenta *Watson Knowledge Studio*

3.6 Teste Arquitetura Extração de Dados

Por forma a avaliar se é possível retirar informações utilizando o artefacto criado, foi escolhido um excerto de comentários do *dataset* e uma imagem que ilustre o modelo de carro para servir de teste.

O objetivo deste ponto passa por perceber se a arquitetura criada tem a capacidade de:

- classificar a imagem do carro e associar a uma categoria de carros; e
- encontrar um conjunto de relações, palavras-chave, entidades, o sentimento, e emoção no respetivo excerto de comentários;

Na aplicação *Postman*, para poder proceder ao pedido de classificar a imagem do carro "dodge_pickup_3500", é necessário definir como o tipo de pedido "*POST*", e no *URL* passar o seguinte *link* "https://gateway-a.watsonplatform.net/visual-

recognition/api/v3/classify?api_key=7036d724f5001570309ee85275c1d46919e9a39c&version=2016-05-20".

Neste *URL* é passado uma *api_key* onde se refere ao serviço criado no *Visual Recognision* da *IBM Bluemix* para ser utilizado.

De seguida, foi preciso definir os campos "*image_file*" onde foi inserido a imagem que vai ser classificada, e foi definido o campo "*classifier_ids*" onde vai ser inserido a chave do classificador customizado.

O aspeto do pedido é ilustrado na figura 11.

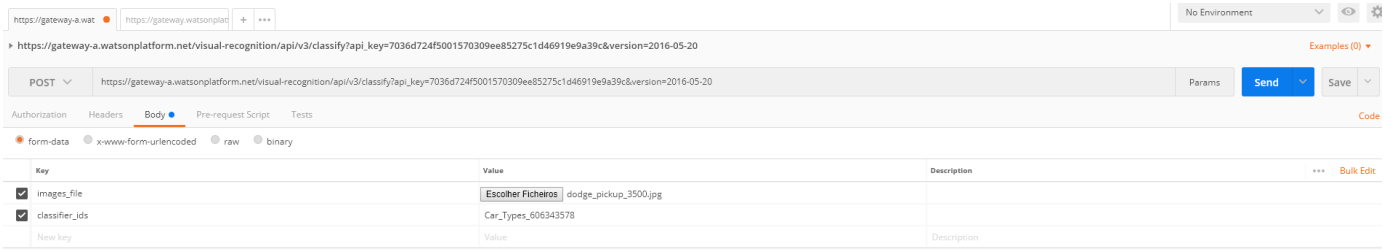


Figura 11 - Pedido Postman

Retirado da ferramenta Postman

Após o pedido estar preparado, foi enviado o pedido, e foi obtido o seguinte resultado apresentado na figura 12.


```
{
    "custom_classes": 19,
    "images": [
        {
            "classifiers": [
                {
                    "classes": [
                        {
                            "class": "Coupe",
                            "score": 0.569759
                        },
                        {
                            "class": "Pickup_truck",
                            "score": 0.998671
                        },
                        {
                            "class": "Sport_utility",
                            "score": 0.84527
                        }
                    ],
                    "classifier_id": "Car_Types_606343578",
                    "name": "Car_Types"
                }
            ],
            "image": "dodge_pickup_3500.jpg"
        }
    ],
    "images_processed": 1
}
```


Figura 12 - Resultado do pedido classificação do Carro

Retirado da ferramenta Postman

Como é ilustrado na figura 12, o classificador de imagens conseguiu obter um *score* com a precisão de 0,998671 a classificação deste carro como sendo da categoria "*Pickup_truck*", sendo que o valor deste *score* varia entre 0 e 1.

Por outro lado, utilizando na mesma a aplicação *Postman*, foi procedido ao pedido de análise dos comentários sobre o carro "*dodge_pickup_3500*".

Para poder realizar o pedido, é necessário escolher o tipo de pedido "*POST*" e no *URL* colocar o seguinte *link* "https://gateway.watsonplatform.net/natural-language-understanding/api/v1/analyze?version=2017-02-27".

De seguida, procede-se no cabeçalho a identificação da autenticação, onde é obtida no serviço *Natural Language Understanding*.

O formato da autenticação e do tipo de pedido é ilustrado na figura 13.

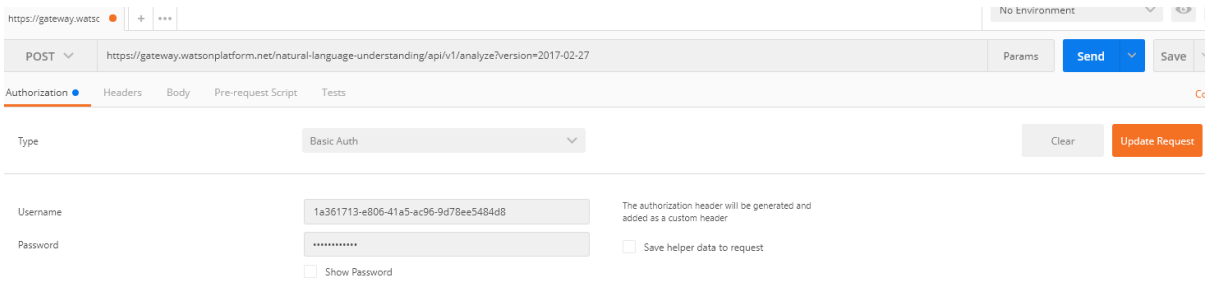


Figura 13 - Formato autenticação

Retirado da ferramenta Postman

O próximo passo passa por escolher no cabeçalho o formato da resposta, utilizando o campo "*Content-Type*" e como atributo "*application/json*".

Depois deste ponto estar definido, o último passo antes de se proceder ao pedido consiste em aceder ao campo "*Body*", escolher o formato "*Raw*" e nessa caixa de texto definir um conjunto de parâmetros: o texto que se pretende analisar, assim como o que se pretende retirar desse texto utilizando a ferramenta *Natural Language Understanding*.

O formato do *Body* é apresentado na figura 14.

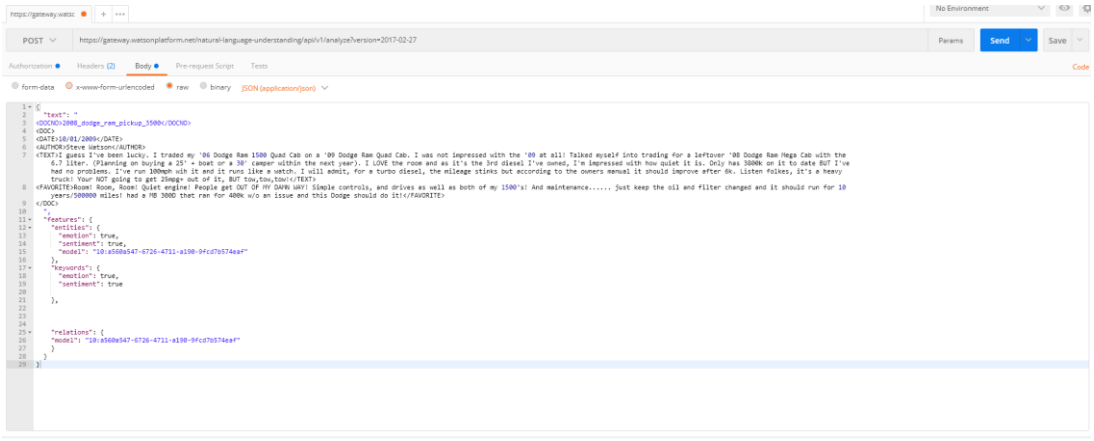


Figura 14 - Formato Body Postman

Retirado da ferramenta Postman

Como podemos verificar na figura 14, o pedido mostra o texto que se pretende analisar, e de seguida na parte dos "*features*" vai ser escolhido os pontos que se vão utilizar para analisar o texto.

No ponto das *entities* vai ser procurado as entidades referidas no modelo customizado, e associado a essas entidades, vai ser atribuído um *score* para avaliar a emoção e o sentimento.

No ponto *keywords*, vai ser procurado palavras-chaves utilizando o modelo por defeito, e associado a essas palavras-chaves vai ser atribuído um *score* para avaliar a emoção e sentimento.

No ponto *relations*, irá ser identificado um conjunto de relações definidas no modelo customizado e atribuído um score para avaliar a precisão.

O próximo passo foi a realização do pedido, no qual o resultado foi o seguinte, ilustrado nas figuras 15, 16 e 17.

```
"keywords": [
    {
        "text": "Dodge Ram",
        "sentiment": {
            "score": -0.386573
        },
        "relevance": 0.983829,
        "emotion": {
            "sadness": 0.154649,
            "joy": 0.327314,
            "fear": 0.081668,
            "disgust": 0.089378,
            "anger": 0.162617
        }
    },
    {
        "text": "Dodge Ram Quad",
        "sentiment": {
            "score": 0
        },
        "relevance": 0.850009,
        "emotion": {
            "sadness": 0.159338,
            "joy": 0.33198,
            "fear": 0.083904,
            "disgust": 0.077779,
            "anger": 0.129812
        }
    },
    {
        "text": "Dodge Ram Mega",
        "sentiment": {
            "score": -0.386573
        },
        "relevance": 0.832531,
        "emotion": {
            "sadness": 0.16824,
            "joy": 0.18308,
            "fear": 0.10322,
            "disgust": 0.088376,
            "anger": 0.141045
        }
    },
```


Figura 15 - Resultado Keywords

Retirado da ferramenta Postman


```
"entities": [
    {
        "type": "Car_Noise",
        "text": "quiet",
        "sentiment": {
            "score": 0.834778
        },
        "emotion": {
            "sadness": 0.06872,
            "joy": 0.320096,
            "fear": 0.093837,
            "disgust": 0.043392,
            "anger": 0.118807
        },
        "disambiguation": {
            "subtype": [
                "NONE"
            ]
        },
        "count": 1
    },
    {
        "type": "Good",
        "text": "Quiet",
        "sentiment": {
            "score": 0.746373
        },
        "emotion": {
            "sadness": 0.031775,
            "joy": 0.294631,
            "fear": 0.093196,
            "disgust": 0.017247,
            "anger": 0.150344
        },
        "disambiguation": {
            "subtype": [
                "NONE"
            ]
        },
        "count": 1
    },
```


Figura 16 - Resultado entities

Retirado da ferramenta Postman

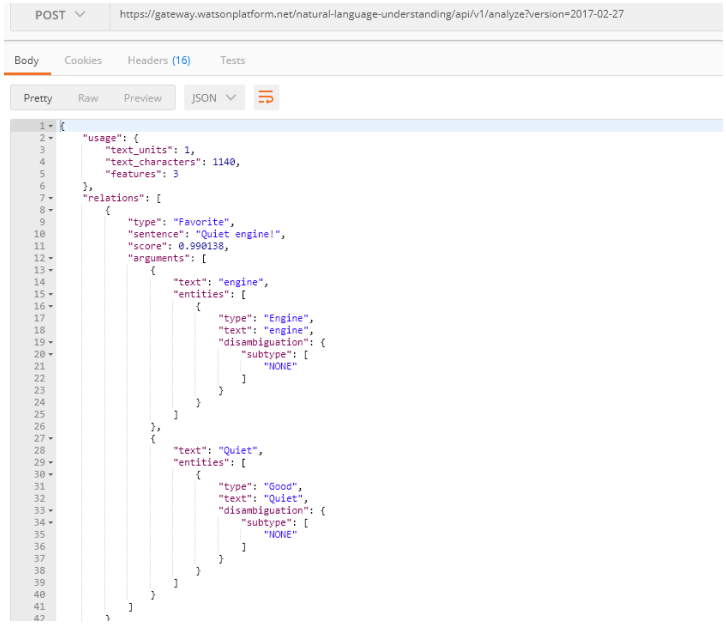


Figura 17 - Resultado Relations

Retirado da ferramenta Postman

Como foi ilustrado na imagem 16, o modelo foi capaz de encontrar 3 entidades distintas, e tal como nas palavras-chave, presente na figura 15, foi associado a cada entidade o sentimento associado, se foi negativo ou positivo, através da variável "*score*" e também foi associado a emoção respetiva através do *score* das variáveis "*Emotion_sadness*", "*Emotion_Joy*", "*Emotion_fear*", "*Emotion_disgust*" e "*Emotion_anger*".

Na figura 17, a ferramenta *Natural Language Understanding* conseguiu encontrar uma relação "*Favorite*" com duas entidades distintas, nomeadamente "*Engine*" e "*Quiet*".

A relação "*Favorite*" com as duas entidades foram identificadas pelo modelo e obteve um *score* de previsão de 0,990138 no qual é bastante satisfatória sabendo que o score tem como valor entre 0 e 1.

Em relação as palavras-chave, a tabela 17 demonstra parte dos resultados obtidos.

Tabela 17 - Resultado Palavras Chave

Text	Sentiment_score	Relevance	Emotion_sadness	Emotion_joy	Emotion_fear	Emotion_disgust	Emotion_anger
Dodge Ram	-0.386573	0.983829	0.154649	0.327314	0.081668	0.089378	0.162617
Dodge Ram Quad	0	0.850009	0.159338	0.33198	0.083904	0.077779	0.129812
Dodge Ram Mega	-0.386573	0.832531	0.16824	0.18308	0.10322	0.088376	0.141045
Quad Cab	0	0.788727	0.159338	0.33198	0.083904	0.077779	0.129812
drives	0.689483	0.431734	0.041541	0.466657	0.131765	0.027324	0.124281
mileage	-0.409288	0.441477	0.125106	0.00742	0.181092	0.751959	0.232211

Na tabela 17, podemos encontrar associado à palavra-chave o "*sentiment_score*", no qual vai avaliar se o sentimento associado a essa palavra-chave é positivo (caso seja maior que 0) ou negativo, caso seja menor que 0.

Em relação as colunas "*Emotion_sadness*", "*Emotion_Joy*", "*Emotion_fear*", "*Emotion_disgust*" e "*Emotion_anger*" vai ser medido a emoção associada a essa palavra-chave. Este valor varia entre 0 e 1.

3.7 Plataforma analítica

Neste ponto irá ser mostrado o processo de criação da plataforma analítica, utilizando Cubos *OLAP* para proceder a uma análise dos dados obtidos.

Na figura 18 está ilustrado os passos até a construção dos cubos.

Figura 18 - Construção dos cubos

Como esta ilustrado na figura 18, os dados obtidos do ponto de extração de dados, explicado no ponto 3.4, são apresentados na ferramenta *Postman*. Esta ferramenta, guarda os ficheiros no formato *Json*. Para poder inserir estes tipos de dados no *SQL Server*, foi necessário recorrer a ferramenta *Talend*, onde foi realizado tarefas de *ETL*, e de seguida, inserir os dados numa base de dados de estágio na ferramenta *SQL Server*. Para se proceder a criação dos cubos, primeiro foi necessário utilizar *o SSIS* para utilizar a função de *ETL* de forma a efetuar um conjunto de transformações para os dados ficarem nos formatos que se pretendiam para inserir nos cubos. Depois de realizar as transformações *ETL* no *SSIS,* o próximo passou por inserir os dados num modelo multidimensional. Para tal, foram criados três projetos Cubos *OLAP*. Foi necessário criar 3 cubos *OLAP* distintos, pois as *entities*, *relations* e *keywords* estudavam informações diferentes e não relacionadas. As criações dos cubos seguiram as recomendações de Kimball & Ross (2011), onde indica que a criação de um modelo multidimensional deve-se focar em eventos mensuráveis, dividindo os dados em medidas ou em contexto descritivo respondendo a questões como "Quem?", "O quê?" "Onde?", "Quando", "Porquê?" e "Como?". Os cubos *OLAP* criados, contem um conjunto de dimensões com dados descritivos e estão conectados numa tabela de factos onde vai possuir as chaves estrangeiras das dimensões assim como um conjunto de medidas. A criação de cubos *OLAP* teve como principal razão o facto de ser flexível quanto às analises que se pretendam realizar, e também é possível realizar uma análise utilizando as medidas sob diferentes perspetivas (dimensões).

De seguida é apresentado um conjunto de imagens a ilustrar os cubos criados, as hierarquias e cálculos associados aos cubos.

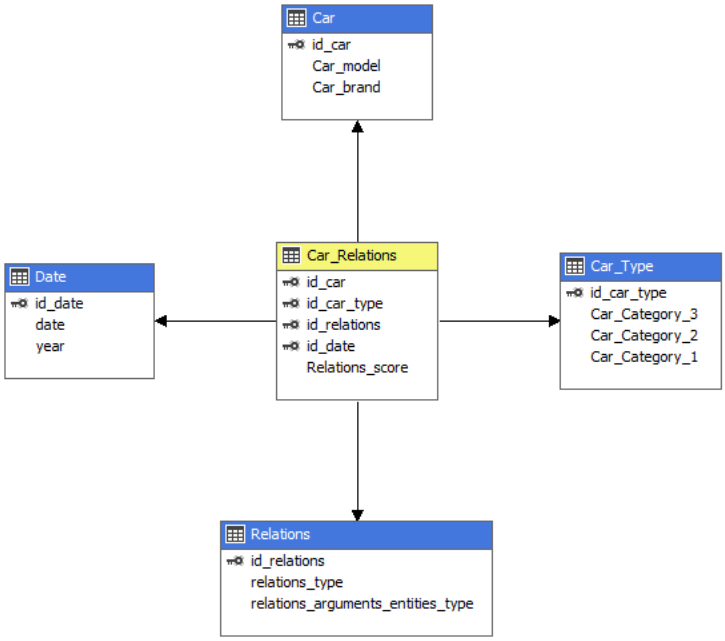


Figura 19 - Cubo Relations

Retirado da ferramenta SSAS

Para o cubo *Relations*, foi adicionado apenas uma medida sendo ela "*Relations_score*".

Como dimensões foram adicionados a "*Date*" (contendo o ano do modelo do carro), "*Car*", contendo o modelo do carro e a marca, o "*Car_type*", contendo a categoria e subcategorias associadas a cada carro, e por último a dimensão "Relations", contendo o tipo de relação e qual o tipo de entidade associado a essa relação.

As Hierarquias criadas foram as seguintes ilustradas nas figuras 20, 21 e 22.

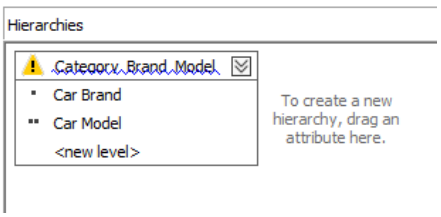


Figura 20 - Relations Hierarquia Carro

Retirado da ferramenta SSAS

A hierarquia "*Category_Brand_Model*" é composta pela marca do carro "*Car_Brand*", seguida do modelo do mesmo "*Car_Model*".

A hierarquia "*Car_Category*" é composta por 3 níveis de categorias de carro "*Car_Category_1*", "*Car_Category_2*" e "*Car_Category_3*".

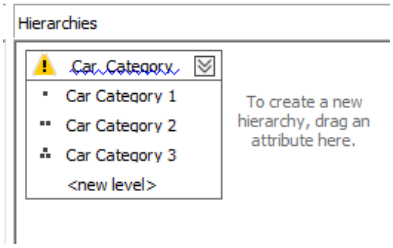


Figura 21 - Relations Categoria carro

Retirado da ferramenta SSAS

A hierarquia "*Date_hierarchy*" é composta pela data "*Date*" e pelo Ano "*Year*".

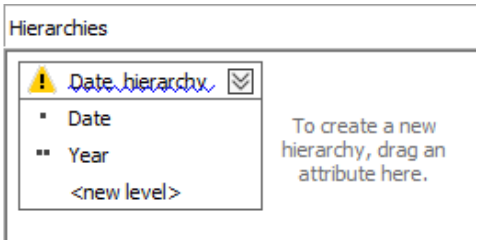


Figura 22 - Relations Hierarquia Data

Retirado da ferramenta SSAS

Como calculo associado ao cubo *Relations*, foi adicionado o seguinte:

- **Median_Score**: SUM([Measures].[Relations Score])/[Measures].[Car Relations Count];

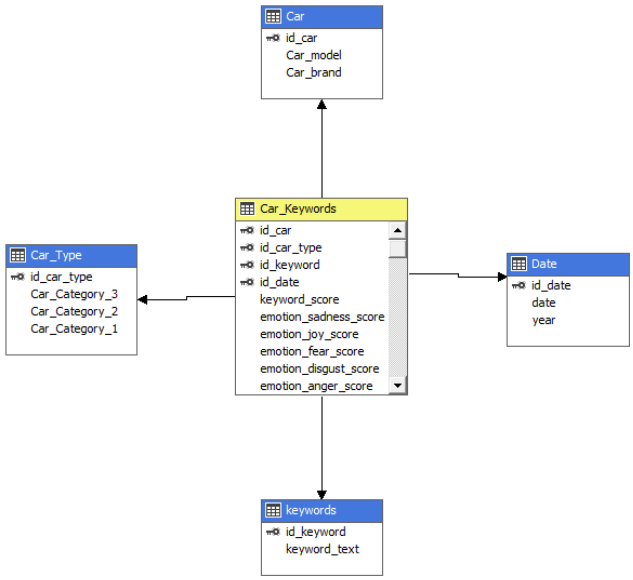


Figura 23 - Cubo Keywords

Retirado da ferramenta SSAS

Para o cubo *Keywords*, foi adicionado um conjunto de medidas, sendo elas: "*keyword_score*", valor numérico para avaliar o sentimento associado à palavra-chave, e os seguintes:" *emotion_sadness_score*", "emotion_*joy_score*", "*emotion_fear_score*", "*emotion_disgust_score*" e "*emotion_anger_score*", para avaliar a emoção associada à palavra-chave.

As dimensões associadas ao cubo Keywords foram as seguintes: "Car", contendo o modelo e a marca do carro, "Car_Type", contendo a categoria e subcategorias associadas ao carro, "Date", contendo o ano e a data, e por último "Keywords", contendo a palavra-chave que se pretende estudar.

As Hierarquias criadas foram as seguintes ilustradas nas figuras 24, 25 e 26.

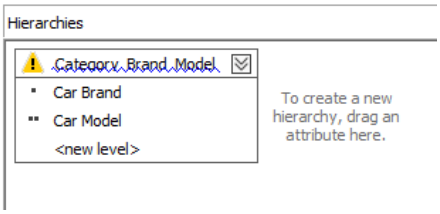


Figura 24 - Keywords Hierarquia Carro

Retirado da ferramenta SSAS

A hierarquia "*Category_Brand_Model*" é composta pela marca do carro "*Car_Brand*", seguida do modelo do mesmo "*Car_Model*".

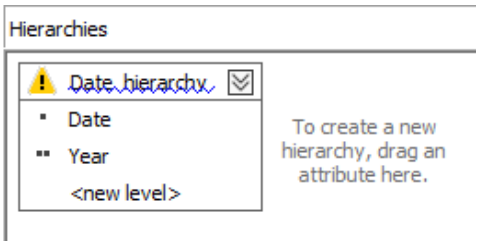


Figura 25 - Keywords Categoria Data

Retirado da ferramenta SSAS

A hierarquia "*Date_hierarchy*" é composta pela data "*Date*" e pelo Ano "*Year*".

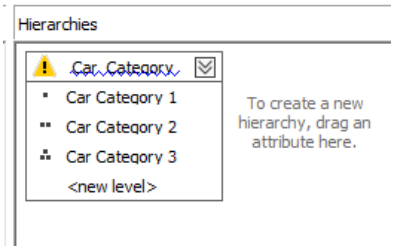


Figura 26 - Keywords Categoria Carro

Retirado da ferramenta SSAS

A hierarquia "*Car_Category*" é composta por 3 níveis de categorias de carro "*Car_Category_1*", "*Car_Category_2*" e "*Car_Category_3*".

Como cálculos associados ao cubo Keywords, foram adicionados os seguintes:

- **Median_score:** SUM([Measures].[Keyword Score])/[Measures].[Car Keywords Count];
- **Median_anger:** SUM([Measures].[Emotion Anger Score])/[Measures].[Car Keywords Count];
- **Median_sadness:** SUM([Measures].[Emotion Sadness Score])/[Measures].[Car Keywords Count];
- **Median_disgust:** SUM([Measures].[Emotion Disgust Score])/[Measures].[Car Keywords Count];
- **Median_fear:** SUM([Measures].[Emotion Fear Score])/[Measures].[Car Keywords Count];

- **Median_joy:** SUM([Measures].[Emotion Joy Score])/[Measures].[Car Keywords Count];

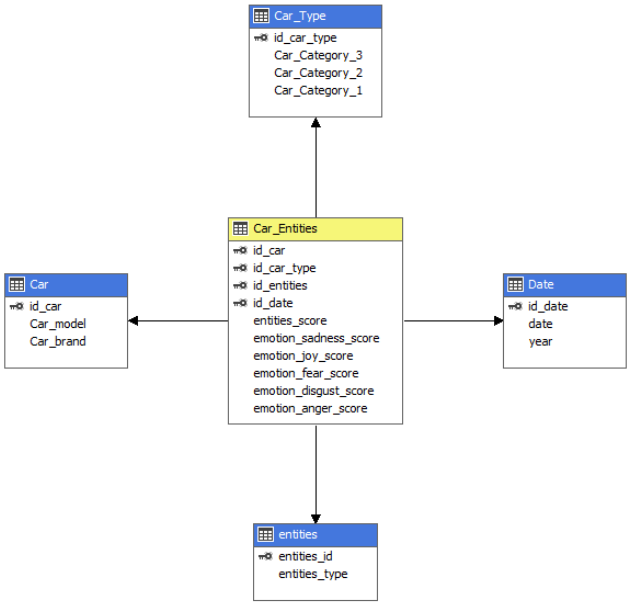


Figura 27 - Cubo Entities

Retirado da ferramenta SSAS

No cubo *Entities*, foi adicionado um conjunto de medidas, sendo elas as seguintes: "*entities score*", valor numérico para avaliar o sentimento associado a uma entidade, e os seguintes: *emotion_sadness_score*", "*emotion_joy_score*", "*emotion_fear_score*", "*emotion_disgust_score*" e "*emotion_anger_score*", para avaliar a emoção associada a uma entidade.

As dimensões associadas ao cubo *entities* foram as seguintes: "*Car*", contendo o modelo e a marca do carro, "*Car_Type*", contendo a categoria e subcategorias associadas ao carro, "*Date*", contendo o ano e a data e por ultimo "*entities*", contendo a entidade que se pretende estudar.

As Hierarquias criadas foram as seguintes ilustradas nas figuras 28, 29 e 30.

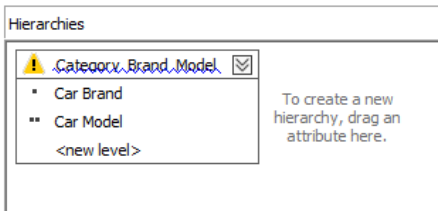


Figura 28 - Entities hierarquia Carro

Retirado da ferramenta SSAS

A hierarquia "*Category_Brand_Model*" é composta pela marca do carro "*Car_Brand*", seguida do modelo do mesmo "*Car_Model*".

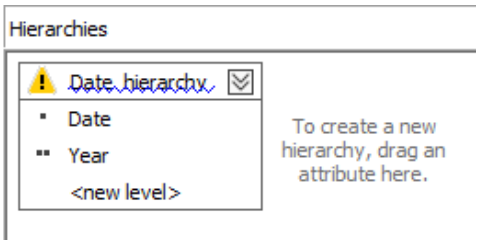


Figura 29 - Entities hierarquia Data

Retirado da ferramenta SSAS

A hierarquia "*Date_hierarchy*" é composta pela data "*Date*" e pelo Ano "*Year*".

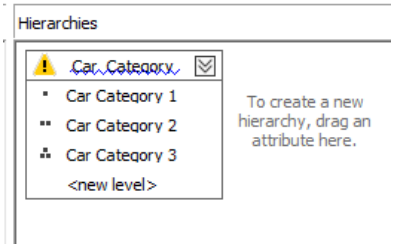


Figura 30 - Entities Categoria Tipo de Carro

Retirado da ferramenta SSAS

A hierarquia "*Car_Category*" é composta por 3 níveis de categorias de carro "*Car_Category_1*", "*Car_Category_2*" e "*Car_Category_3*".

Como cálculos associados ao cubo *Entities*, foram adicionados os seguintes:

- **Median_score:** SUM([Measures].[Entities Score])/[Measures].[Car Entities Count];

- **Median_joy:** SUM([Measures].[Emotion Joy Score])/[Measures].[Car Entities Count];
- **Median_anger:** SUM([Measures].[Emotion Anger Score])/[Measures].[Car Entities Count];
- **Median_disgust:** SUM([Measures].[Emotion Disgust Score])/[Measures].[Car Entities Count];
- **Median_fear:** SUM([Measures].[Emotion Fear Score])/[Measures].[Car Entities Count];
- **Median_sadness:** SUM([Measures].[Emotion Sadness Score])/[Measures].[Car Entities Count];

4 Discussão dos Resultados

As análises ilustradas neste capítulo foram realizadas recorrendo a ferramenta Excel. Foram utilizados os cubos gerados no ponto 3.7 e procedeu-se a sua análise com o objetivo de perceber o tipo de informações que foi possível obter do seu estudo. Foram realizadas três análises distintas: uma análise geral ilustrando a análise de sentimento associada aos carros entre outros componentes utilizando os três cubos *OLAP*, uma análise de emoção dos diversos carros associados a categoria "*Pickup_truck*" utilizando o cubo *Entities*; e por último, uma análise detalhada do modelo de carro "s40" da marca "volvo", com a análise de sentimento, quais os componentes bons e maus, usando os três cubos *OLAP*.

Na figura 31 é apresentado uma visão global dos cubos no Excel.

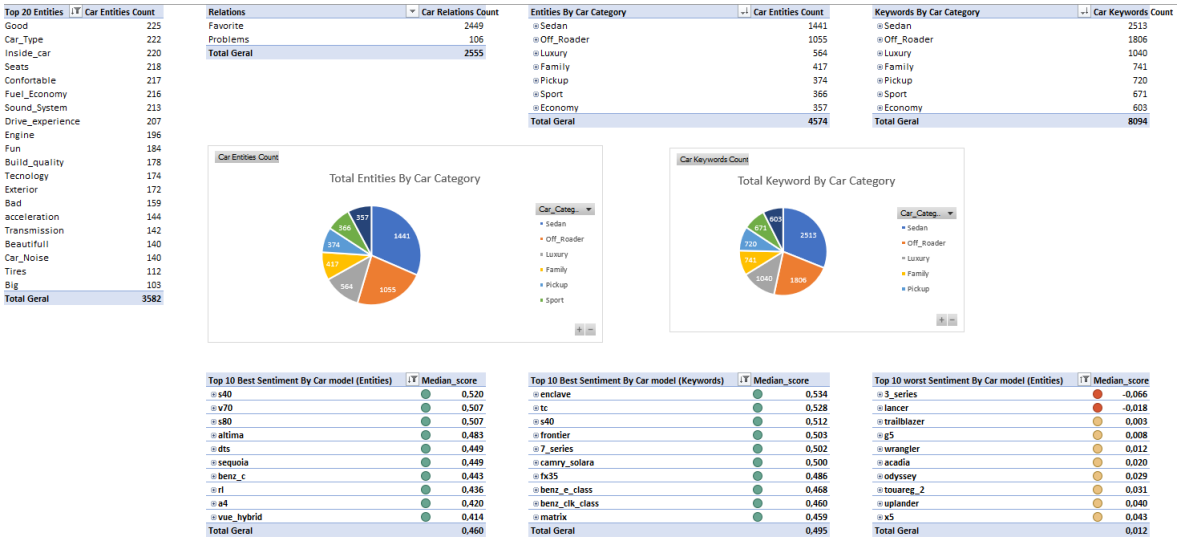


Figura 31 - Análise Visão Global

Retirado da ferramenta Excel

Na figura 31, no canto superior esquerdo, são apresentadas as 20 *Entities* mais utilizadas nos comentários sobre os carros. É possível perceber que a entidade mais referenciada foi "*Good*" estando, portanto, relacionado a algum componente positivo. É facilmente percetível que as entidades mais referenciadas são sobre o tipo de carro, a parte interior do carro, os assentos, o facto de ser confortável, o facto de ser um carro económico (gasto de combustível), entre outros.

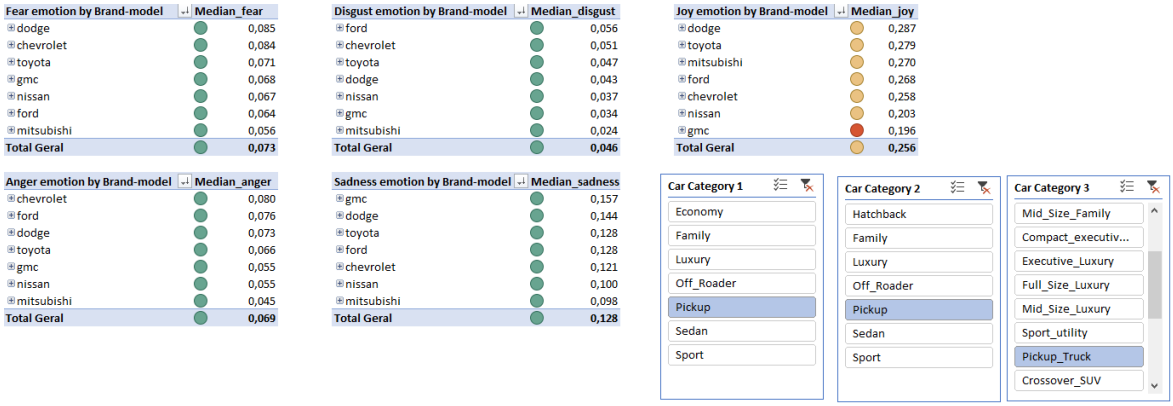
Na seguinte tabela à direita, é apresentado o número de relações do tipo *favorite* e *problems*, no qual se percebe que o sistema montado foi capaz de encontrar mais facilmente a relação de *favorite* do que *problems*. De seguida, nas duas tabelas à direita são apresentadas o número de entidades e palavras-chave por categoria de carros. Percebe-se que existe uma categoria que

apresenta maior número de referencias (categoria *Sedan*), a segunda maior é a categoria *Off_roader*, e as seguintes são distribuídas mais ou menos da mesma forma.

Nas duas figuras abaixo, estão apresentadas as duas tabelas em cima referenciadas, mas num formato diferente, num gráfico circular divido pelas categorias. Nas tabelas em baixo, no lado esquerdo e a do meio (utilizando o cubo *entities* e *keywords*), são apresentados os 10 melhores modelos de carro utilizando a medida "*median_score*" que mede a análise de sentimento associado aos carros através da média do sentimento associado as *entities* e *keywords*. Ao comparar estas duas tabelas, percebe-se que existem bastantes diferenças, mas existe um modelo de carro que está bem cotado nas duas tabelas (o modelo s40). Esta grande diferença é devida ao facto que na identificação das palavras-chaves foi utilizado o modelo de *data mining* de anotações por defeito da ferramenta, e para a identificação das entidades, foi utilizada um modelo de *data mining* de anotações customizado.

No canto inferior direito é apresentado uma tabela (utilizando o cubo *Entities*) contendo os 10 piores carros, utilizando como medida a "median_score" que mede o sentimento através da média. É facilmente percetível que os dois piores carros são os modelos "serie_3" e "lancer", utilizando a formatação condicional do Excel, foi possível distinguir os valores negativos (com uma bola vermelha), dos valores neutros (bola a amarelo).

De seguida, é apresentado uma figura a ilustrar a análise de emoção dos comentários sobre os carros.



Fear emotion by Brand-model	Median_fear
dodge	0,085
chevrolet	0,084
toyota	0,071
gmc	0,068
nissan	0,067
ford	0,064
mitsubishi	0,056
Total Geral	**0,073**

Disgust emotion by Brand-model	Median_disgust
ford	0,056
chevrolet	0,051
toyota	0,047
dodge	0,043
nissan	0,037
gmc	0,034
mitsubishi	0,024
Total Geral	**0,046**

Joy emotion by Brand-model	Median_joy
dodge	0,287
toyota	0,279
mitsubishi	0,270
ford	0,268
chevrolet	0,258
nissan	0,203
gmc	0,196
Total Geral	**0,256**

Anger emotion by Brand-model	Median_anger
chevrolet	0,080
ford	0,076
dodge	0,073
toyota	0,066
gmc	0,055
nissan	0,055
mitsubishi	0,045
Total Geral	**0,069**

Sadness emotion by Brand-model	Median_sadness
gmc	0,157
dodge	0,144
toyota	0,128
ford	0,128
chevrolet	0,121
nissan	0,100
mitsubishi	0,098
Total Geral	**0,128**

Figura 32 - Análise de emoção

Retirado da ferramenta Excel

Na figura 32 são apresentadas cinco tipo de emoções distintas associadas a um conjunto de modelos de carros. A escolha destes modelos de carros associados a estas emoções, deve-se ao

facto de ter sido criado uma segmentação de dados apresentados como "*Car Category 1*", "*Car Category 2*" e "*Car Category 3*", onde foi escolhido apenas os modelos de carros que pertenciam à "Car Category 1" "*Pickup*", "*Car Category 2*" "*Pickup*" e "*Car Category 3*" "*Pickup_Truck*". Para cada uma das 5 tabelas, foi utilizado a medida média para medir a emoção do modelo e carro, e também foi adicionado uma formatação condicional para perceber se o valor é aceitável, neutro ou mau. As tabelas que medem as emoções "*anger*", "*sadness*", "*fear*" e "*disgust*" não possuem valores com grande impacto, o que significa que não se associam a estas quatro emoções. A emoção *joy*, possui um maior impacto que as outras, mas mesmo assim, a maioria dos modelos possui um valor considerado neutro, apenas o modelo "*gmc*" ficou abaixo do normal.

Por ultimo, é apresentado uma figura a ilustrar uma possível análise de vários fatores a um modelo de carro especifico.

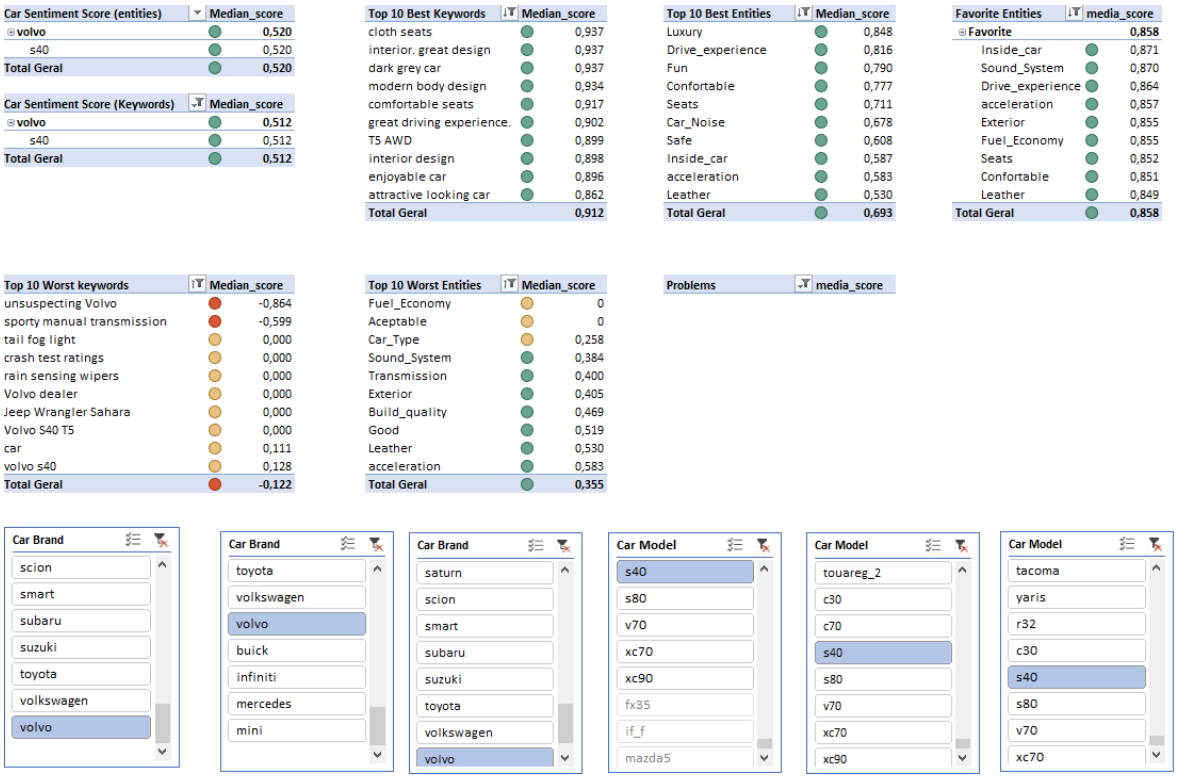


Figura 33 - Análise geral carro

Retirado da ferramenta Excel

Na figura 33 é apresentado um conjunto de tabelas com diferentes fatores, associado a um modelo de carro da marca "*volvo*" e modelo "*s40*". A escolha da análise deste modelo específico, deveu-se ao facto de na figura 31, ter sido considerado um dos melhores carros utilizando a variável "*median_score*", que mede através da média o sentimento dos comentários. Nesta análise, foi

utilizado a segmentação da hierarquia "*Category_Brand_model*" dos 3 cubos (*Relations, Keywords e Entities*).

No canto superior esquerdo, é mostrado o *score* obtido da análise de sentimento do modelo de carro específico, quer do cubo *Entities*, como do cubo *Keywords*. Consiste num valor bastante bom, o quer dizer que as pessoas ficaram com uma boa impressão do carro.

Nas três tabelas a direita, é mostrado o top 10 *keywords* (do cubo *Keywords*), top 10 *entities* (do cubo *Entities*) e top 10 *Favorite entities* (do cubo *Relations*), todas estas utilizando uma medida que calcula a média. Na tabela top 10 *keywords*, existe uma boa referência dos assentos, do interior do carro, o que revela ser um carro com alguns pontos fortes no conforto do interior do carro. Na tabela top 10 *entities*, foi identificado como positivo o facto de ser um carro de luxo, o facto de ser confortável, o barulho que o carro produz, o facto de ser um carro seguro. Na tabela *Favorite entities*, foi identificado como pontos fortes o interior do carro, os assentos, o sistema de som, a experiência de condução, entre outros.

Nas três tabelas em baixo, é ilustrado o "*Top 10 Wrost Keywords*" (do cubo *Keywords*), "*Top 10 Wrost Entities*" (do cubo *Entities*) e "*Problems*" (do cubo *Relations*) utilizando como medida o "*median_score*" que mede o sentimento através da média. Na tabela "*Top 10 Wrost Keywords*", apenas apresenta duas palavras-chaves como sendo um ponto negativo, e as restantes são componentes neutras, o que revela a pouca quantidade de problemas encontrados neste modelo de carro. Na tabela "*Top 10 Entities*", não foi encontrado nenhum componente como sendo negativo, apenas dois como sendo neutros e os restantes sendo positivos. Na ultima tabela, "*Problems*", não foi encontrado qualquer problema identificado nos comentários.

Face estas diferentes análises apresentadas em cima, os resultados obtidos foram bastantes satisfatórios, pois foi possível retirar informações relevantes acerca dos carros como: sentimento e emoção associado a diversos componentes dos carros (positivo e negativo); Quais os problemas/pontos fracos e componentes que as pessoas gostaram mais/pontos fortes; Quais os melhores/piores carros; permitiu realizar uma análise geral dos carros assim como uma análise mais pormenorizada acerca de um modelo de carro.

5 Conclusões, Limitações e Trabalho Futuro

5.1 Conclusões

O levantamento da Revisão de Literatura, permitiu perceber um conjunto de informações relevantes para o enquadramento deste documento. Foi definido o que são dados, os diferentes tipos de dados, assim como os diferentes tipos de análises de dados existentes. Também foi definido uma abordagem histórica de como evoluíram as técnicas de análise de dados ao longo de tempo. Para perceber que técnicas de análise existem para diferentes tipos de dados, foi feito um levantamento de técnicas incluídas em artigos apresentados em conferências científicas, e posteriormente agrupadas pelos diferentes formatos de dados. Este levantamento permitiu perceber que técnicas de análise de dados estavam a ser utilizadas pela comunidade científica para diferentes formatos de dados.

No que diz respeito a uma visão mais tecnológica, foi procedido ao levantamento de ferramentas de análise para explorar diferentes formatos de dados.

Posteriormente, procedeu-se à criação do artefacto que permitiu responder ao problema proposto nesta dissertação, definido no ponto 1.3.1. Percebeu-se rapidamente, que para tirar partido de dados não estruturados e semiestruturados, era necessário encontrar uma forma de os transformar e armazenar. Para tal, foi definido um artefacto composto por uma arquitetura com o objetivo de extrair informações úteis de dados não estruturados e semiestruturados. De seguida, o artefacto foi avaliado e testado, o que permitiu perceber que o artefacto criado é capaz de extrair informações relevantes de dados não estruturados e semiestruturados.

Para enriquecer o artefacto, foi elaborada uma plataforma analítica de forma a proceder a diferentes análises dos dados obtidos da experiência. Na plataforma analítica, foi possível verificar um conjunto de informações relevantes como: a análise de sentimento e emoção; quais os problemas/pontos fortes associados a um carro; quais os melhores/piores carros; permitiu realizar uma análise geral, assim como uma análise pormenorizada dos carros.

Os resultados obtidos foram bastante satisfatórios, visto que todos os objetivos propostos foram alcançados com sucesso, tendo ainda sido publicado o trabalho realizado, através de um poster na conferência CAPSI 2017 em anexo.

5.2 Limitações

Nesta dissertação existiram algumas limitações sendo elas:

- ferramentas com licenças limitadas por tempo nomeadamente *Talend* 30 dias e para as ferramentas da *IBM* 6 meses;
- o tempo gasto à procura das ferramentas com as funcionalidades pretendidas;
- devido a limitações de ordem temporal, não foi possível realizar mais experiências de forma a explorar mais técnicas;
- o artefacto criado, não pode ser utilizado como um sistema genérico. Para cada caso específico, é necessário calibrar as ferramentas de forma a procurar o tipo de informações mais relevantes para o caso de estudo específico, pois, o objeto de estudo pode ser completamente diferente.

5.3 Trabalho Futuro

Como trabalho futuro, foram encontradas diversas oportunidades, sendo elas:

- poder realizar um caso semelhante utilizando dados não estruturados e semiestruturados numa organização real, com vista a mostrar que a variedade de dados pode relevar informações úteis para análise;
- explorar outras técnicas de análise de dados, de forma a perceber que tipo de informações se pode retirar;

Referências

Ackoff, R. L. (1989). From data to wisdom. *Journal of applied systems analysis*, 3-9.

Alshareef, A. M., Bakar, A. A., Hamdan, A. R., Abdullah, S. M., & Alweshah, M. (2015). A case-based reasoning approach for pattern detection in Malaysia rainfall data. *International Journal of Big Data Intelligence*, 285-302.

Anagnostopoulos, C., & Triantafillou, P. (2014). Scaling out big data missing value imputations: pythia vs. godzilla. *In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining* , 651-660.

Asai, T., Abe, K., Kawasoe, S., A., H., Sakamoto, H., & Arikawa, S. (2004). Efficient substructure discovery from large semi-structured data. *IEICE TRANSACTIONS on Information and Systems*, 2754-2763.

Baars, H., & Kemper, H.-G. (2008). Management support with structured and unstructured data—an integrated business intelligence framework. . *Information Systems Management*, 132-148.

Barabási, A. L., & Albert, R. (1999). Emergence of scaling in random networks. *science*, 509-512.

Bellinger, G., Castro, D., & Mills, A. (2004). *Data, information, knowledge, and wisdom.* Obtido de geoffreyanderson.net: geoffreyanderson.net

Bi, B., Ma, H., Hsu, B. J., Chu, W., Wang, K., & Cho, J. (2015). Learning to recommend related entities to search users. *In Proceedings of the Eighth ACM International Conference on Web Search and Data Mining* , (pp. 139-148.

Bitterer, A. (2011). Hype Cycle for Business Intelligence. *Gartner RAS Core Research Note G*.

Blanco, R., Ottaviano, G., & Meij, E. (2015). Fast and space-efficient entity linking for queries. *In Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pp. 179-188.

Borisyuk, F., Kenthapadi, K., Stein, D., & Zhao, B. (2016). CaSMoS: A Framework for Learning Candidate Selection Models over Structured Queries and Documents. *In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 441-450.

Caballero Barajas, K. L., & Akella, R. (2015). Dynamically Modeling Patient's Health State from Electronic Medical Records: A Time Series Approach. *In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 69-78.

Chaudhuri, S., Dayal, U., & Narasayya, V. (2011). An Overview of Business Intelligence Technology. *Communications of the ACM*, 88-98.

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS quarterly*.

Chen, N., Hoi, S. C., Li, S., & Xiao, X. (2015). SimApp: A framework for detecting similar mobile applications by online kernel learning. *In Proceedings of the Eighth ACM International Conference on Web Search and Data Mining* , pp. 305-314.

Chen, S., & Joachims, T. (2016). Predicting Matchups and Preferences in Context. *KDD*, 775-784.

Chen, Z., & Liu, B. (2014). Mining topics in documents: standing on the shoulders of big data. . *In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 1116-1125.

Chu, L., Wang, Z., Pei, J., Wang, J., Zhao, Z., & Chen, E. (2016). Finding Gangs in War from Signed Networks., (pp. 1505-1514).

Cukier, K. (2010). Data everywhere: A special report on managing information. *Economist Newspaper*.

Davenport, T. H., & Prusak, L. (1998). *Working knowledge: How organizations manage what they know.*

Dimitrakopoulos, G., Chatzigiannakis, V., & Tsitouras, L. (2017). A knowledge-based integrated framework for increasing social management intelligence . *International Journal of Big Data Intelligence*, 36-46.

Dong, X., Gabrilovich, E., Heitz, G., Horn, W., Lao, N., Murphy, K., & Zhang, W. (2014). Knowledge vault: A web-scale approach to probabilistic knowledge fusion. *In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 601-610.

Evans, J. R., & Lindner, C. H. (2012). Business analytics: the next frontier for decision sciences. *Decision Line*, 4-6.

Ganesan, K., & Zhai, C. (2012). Opinion-based entity ranking. *Information retrieval*, 116-150.

Gao, S., Ma, J., & Chen, Z. (2015). Modeling and predicting retweeting dynamics on microblogging platforms. *In Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pp. 107-116.

Geerdink, B. (2015). A reference architecture for big data solutions-introducing a model to perform predictive analytics using big data technology. . *International Journal of Big Data Intelligence*, 236-249.

Giusto, D., Iera, A., Morabito, G., & Atzori, L. (2010). *The internet of things: 20th Tyrrhenian workshop on digital communications. .* Springer Science & Business Media.

Grabocka, J., Schilling, N., Wistuba, M., & Schmidt-Thieme, L. (2014). Learning time-series shapelets. . *In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 392-401.

Gupta, V., & Lehal, G. S. (2009). A survey of text mining techniques and applications. *Journal of emerging technologies in web intelligence*, 60-76.

Halper, F., & Krishnan, K. (2013). TDWI Big Data Maturity Model Guide Interpreting Your Assessment Score. *TDWI Benchmark Guide*.

Hayashi, K. M., Toyoda, M., & Kawarabayashi, K. I. (2015). Real-time top-r topic detection on twitter with topic hijack filtering. *In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 417-426.

Huo, Z., Nie, F., & Huang, H. (2016). Robust and Effective Metric Learning Using Capped Trace Norm: Metric Learning via Capped Trace Norm. *In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1605-1614.

Järvinen, P. (2007). Action research is similar to design science. *Quality & Quantity*, 37-54.

Kıcıman, E., & Richardson, M. (2015). Towards decision support and goal achievement: Identifying action-outcome relationships from social media. . *In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* , 547-556.

Kimball, R., & Ross, M. (2011). *The data warehouse toolkit: the complete guide to dimensional modeling.* John Wiley & Sons.

Kokkodis, M., Papadimitriou, P., & Ipeirotis, P. G. (2015). Hiring behavior models for online labor markets. *In Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pp. 223-232.

Kotzias, D., Denil, M., De Freitas, N., & Smyth, P. (2015). From group to individual labels using deep features. . *In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 597-606.

Kurashima, T., Iwata, T., Takaya, N., & Sawada, H. (2014). Probabilistic latent network visualization: inferring and embedding diffusion networks. . *In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 1236-1245.

Li, L., Yao, Y., Tang, J., Fan, W., & Tong, H. (2016). QUINT: On Query-Specific Optimal Networks. *In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 985-994.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 1-167.

Liu, J., Aggarwal, C., & Han, J. (2015). On integrating network and community discovery. *In Proceedings of the Eighth ACM International Conference on Web Search and Data Mining* , pp. 117-126.

Lomotey, R. K., & Deters, R. (2015). Unstructured data mining: use case for CouchDB. *International Journal of Big Data Intelligence*, 168-182.

Lusch, R., Liu, Y., & Chen, Y. (2010). The phase transition of markets and organizations: The new intelligence and entrepreneurial frontier. . *IEEE Intelligent Systems*, 71-75.

Makrynioti, N., Grivas, A., Sardianos, C., Tsirakis, N., Varlamis, I., Vassalos, V., & Tsantilas, P. (2017). PaloPro: a platform for knowledge extraction from big social data and the news. *International Journal of Big Data Intelligence*, 3-22.

Mu, X., Zhu, F., Zhou, Z. H., Lim, E. P., Xiao, J., & Wang, J. (2016). User Identity Linkage by Latent User Space Modelling. *Proceedings of 22nd ACM SIGKDD international conference on knowledge discovery and data mining*.

Mukherjee, S., Weikum, G., & Danescu-Niculescu-Mizil, C. (2014). People on drugs: credibility of user statements in health communities. *In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining* , 65-74.

Nagarajan, M., Wilkins, A. D., Bachman, B. J., Novikov, I. B., Bao, S., Haas, P. J., & Regenbogen, S. (2015). Predicting future scientific discoveries based on a networked analysis of the past literature. . *In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019-2028.

Nyce, C., & CPCU, A. (2007). Predictive analytics white paper. *American Institute for CPCU*, 9-10.

Oliveira, T. P., Barbar, J. S., & Soares, A. S. (2016). Computer network traffic prediction: a comparison between traditional and deep learning neural networks. *International Journal of Big Data Intelligence*, 28-37.

Peel, L., Larremore, D. B., & Clauset, A. (2016). The ground truth about metadata and community detection in networks. . *arXiv preprint* .

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2008). A Design Science Research Methodology for Information Systems Research. *Journal of management information systems*, 45-77.

Russom, P. (2011). Big data analytics. *TDWI best practices report*, 1-35.

Sallam, R. L., Richardson, J., Hagerty, J., & Hostmann, B. (2011). Magic quadrant for business intelligence platforms. . *Gartner Group, Stamford, CT.* .

Schubert, E., Weiler, M., & Kriegel, H. P. (2014). Signitrend: scalable detection of emerging topics in textual streams by hashed significance thresholds. . *In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 871-880.

Senellart, P., & Blondel, V. D. (2008). Automatic discovery of similarwords. . *In Survey of Text Mining II* , 25-44.

Sharma, S., Tim, U. S., Gadia, S., Wong, J., Shandilya, R., & Peddoju, S. K. (2015). Classification and comparison of NoSQL big data models. *International Journal of Big Data Intelligence*, 201-221.

Shashidhar, V., Pandey, N., & Aggarwal, V. (2015). Spoken english grading: Machine learning with crowd intelligence. . *In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* , 2089-2097.

Sudhof, M., Goméz, A., Maas, A. L., & Potts, C. (2014). Sentiment expression conditioned by affective transitions and social force. *In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 1136-1145.

Tang, J., Chang, S., Aggarwal, C., & Liu, H. (2015). Negative link prediction in social media. *In Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pp. 87-96.

Tran, N. K., Ceroni, A., Kanhabua, N., & Niederée, C. (2015). Back to the past: Supporting interpretations of forgotten stories by time-aware re-contextualization. *In Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, 339-348.

Turban, E., Sharda, R., Aronson, J. E., & King, D. (2008). *Business Intellligence: A Managerial Approach.* Pearson Prentice Hall.

Ulanova, L., Yan, T., Chen, H., Jiang, G., Keogh, E., & Zhang, K. (2015). Efficient Long-Term Degradation Profiling in Time Series for Complex Physical Systems. . *In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2167-2176.

Veeriah, V., Durvasula, R., & Qi, G. J. (2015). Deep learning architecture with dynamically programmed layers for brain connectome prediction. . *In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1205-1214.

Von Alan, R. H., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS quarterly*, 75-105.

Wang, S., Chen, Z., Fei, G., Liu, B., & Emery, S. (2016). Targeted Topic Modeling for Focused Analysis. *KDD*, (pp. 1235-1244).

Watson, H. J., & Wixom, B. H. (2007). The Current State of Business Intelligence. *IEEE Computer*, 96-99.

Wei, Y., Zheng, Y., & Yang, Q. (2016). Transfer knowledge between cities. . *In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* , 1905-1914.

Wlodarczyk, T. W., & Hacker, T. J. (2014). Current trends in predictive analytics of big data. *International Journal of Big Data Intelligence*, 172-180.

Wu, Y., & Ester, M. (2015). Flame: A probabilistic model combining aspect based opinion mining and collaborative filtering. *In Proceedings of the Eighth ACM International Conference on Web Search and Data Mining* , pp. 199-208.

Yang, F., Milosevic, D., & Cao, J. (2017). Optimising column family for OLAP queries in HBase. *International Journal of Big Data Intelligence*, 23-35.

Yao, Y., Tong, H., Xu, F., & Lu, J. (2014). Predicting long-term impact of CQA posts: a comprehensive viewpoint. . *In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining* , 1496-1505.

Ye, W., Goebl, S., Plant, C., & Böhm, C. (2016). FUSE: Full Spectral Clustering. *KDD*, (pp. 1985-1994).

Zalmout, N., & Ghanem, M. M. (2016). Multivariate adaptive community detection in Twitter. . *International Journal of Big Data Intelligence*, 239-249.

Zhang, C., Zhang, K., Yuan, Q., Zhang, L., Hanratty, T., & Han, J. (2016). GMove: Group-Level Mobility Modeling Using Geo-Tagged Social Media. . *In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1305-1314.

Zhao, Y., & Zhu, Q. (2014). Evaluation on crowdsourcing research: Current status and future direction. . *Information Systems Frontiers*, 417-434.

Zhao, Z., Liu, J., & Cox, J. (2014). Safe and efficient screening for sparse support vector machine. . *In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 542-551.

Zhou, Y., Liu, L., & Buttler, D. (2015). Integrating vertex-centric clustering with edge-centric clustering for meta path graph analysis. *In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* , 1563-1572.

Zhuang, H., & Young, J. (2015). Leveraging in-batch annotation bias for crowdsourced active learning. *In Proceedings of the Eighth ACM International Conference on Web Search and Data Mining* , pp. 243-252.

Zuo, Y., Wu, J., Zhang, H., Lin, H., Wang, F., Xu, K., & Xiong, H. (2016). Topic Modeling of Short Texts: A Pseudo-Document View. . *In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* , 2105-2114.

Anexo

Categorias de carros utilizadas na experiência, na tabela 18.

Tabela 18 - Categorias de Carros

Categorias
Crossover_SUV
Full_Size_Sedan
Mid_Size_Sedan
Compact_executive_luxury
Compact_Family
Executive_Luxury
Sports_Car
Pickup_Truck
Mini_Van_Sedan
Sport_utility
Hatchback
Muscle_Car
Coupe
Mid_Size_Luxury
City_car
Convertible
Full_Size_Luxury
Subcompact
Mid_Size_Family

Na tabela 19, é apresentado a lista completa e Entidades utilizadas na experiência.

Tabela 19 - Entidades

Entidades
Suspension
Oil_change
Rims
Weels
Sport
Luxury
Safe
AWD
Downhill_assist
Stability_control
Batteries
Technology
Lights
Radiator
Thermostast

Water_pump
Perfect
Plastic
Beautifull
Aceptable
Confortable
Normal
High
Low
Fun
Small
Big
Replacement
Bad
Good
Tires
Price
Build_quality
Connected_phone_system
DVD
On_board_computer
Exterior
Design
acceleration
Traction_control
Airbag
Cargo_space
Sunroof
Doors
Door_lock
Air_conditioner
Cost_of_car
ABS
Roof
Exhaust_system
Transmission
Car_noise
Parking_Sensors
Car_Type
Drive_experience
Inside_car
Brakes
Fuel_Economy
Seats
Sound_System
Windows

Engine

Está prevista uma publicação de um artigo científico na conferência:

- Worldcist.

Resumo do poster entregue na CAPSI 2017:

Data Analytics para Variedade de Dados

Data Analytics for Data Variety

Tiago Emanuel Senra da Cruz, Universidade do Minho, Portugal, a66785@alunos.uminho.pt

Jorge Oliveira e Sá, Centro ALGORITMI, Universidade do Minho, Portugal, jos@dsi.uminho.pt

Resumo

A internet fez com que os gestores das organizações tivessem acesso a grandes quantidades de dados e esses dados são apresentados em diferentes formatos, em concreto, estruturados, semiestruturados e não estruturados. Esta variedade de dados é essencialmente proveniente das redes sociais, mas não só, também são provenientes de sistemas ciber-físicos. Verifica-se que para os dados estruturados existem técnicas validadas, estudadas e maduras, mas para os outros tipos de dados, ou seja, semiestruturados e não estruturados tal já não se verifica. Neste poster, é apresentado um conjunto de técnicas de análise de dados para os dados semiestruturados e não estruturados, utilizando como principal bibliografia conferências de investigação na área de análise de dados.

Palavras-chave: Análise de dados, Variedade de dados, Tipo de dados

Abstract

Through the Internet, the organizations managers had access to massive amounts of data and these data are presented in different formats, namely, structured, semi-structured and unstructured. These variety of data is essentially generated from social networks, but not only, they also are generated from cyber-physical systems, from machines, sensors, among others. While the structured data has techniques well studied, mature and validated, otherwise the other types of techniques, semi-structured and unstructured, this is no longer true. In this poster, a set of data analysis techniques is presented for the semi-structured and unstructured data by using as main bibliography data analytics conferences.

Keywords: *Data Analytics, Data Variety, Data Types.*

Descrição do trabalho (Póster)

Com o aumento da utilização da internet por parte das pessoas e organizações, a quantidade de informação cresceu exponencialmente. O universo digital apresenta uma enorme diversidade de dados gerados através das redes sociais, onde os utilizadores geram conteúdos diversificados como imagens, vídeos, textos, sites, entre outros (Fan e Gordon, 2014), mas não só, também a Internet of Things (Uckelmann et al., 2011), tornou possível que máquinas pudessem comunicar entre si automaticamente, utilizando sistemas de endereçamento exclusivos, e desta forma, consigam trabalhar em conjunto para atingir um fim comum. Esta diversidade de dados é acompanhada por uma variedade de tipo dados, estruturados, semiestruturados e não estruturados, tanto gerados por pessoas como máquinas e consiste numa característica a ter em consideração, pois a sua análise traz valor para as organizações (Russom, 2011).

Os gestores para poderem tirar proveito destes tipos de dados, nomeadamente dos dados estruturados, semiestruturados e não estruturados, terão que utilizar diferentes técnicas de análise de dados capazes de retirar informações valiosas para os ajudar na tomada de decisão. As técnicas de análise de dados do tipo estruturado estão maduras e validadas pela comunidade científica, mas tal não se verifica para os dados do tipo semiestruturado e não estruturado.

Com o trabalho realizado pretende-se identificar técnicas de análise de dados para os tipos de dados semiestruturados e não estruturados com o objetivo de perceber qual o valor que se pode obter através da sua análise.

Nas seguintes tabelas são ilustradas um conjunto de técnicas de análise de dados retirados das conferências: Conference of Knowledge Discovery and Data Mining dos anos 2014, 2015 e 2016, Conference of Web Search and Data Mining do ano 2015 e por último, International Journal of Big Data Intelligence do ano 2014 a 2017; e estão agrupadas pelo tipo de dados (semiestruturado e Não estruturado).

Para Dados Semiestruturados

Referência	Técnica
(Dong, et al., 2014)	HTML trees (DOM)
(Li et al.2016), (Zhou, Liu, & Buttler, 2015)	Link prediction
(Bi, et al., 2015)	probabilistic Three-way Entity Model (TEM)
(Blanco, Ottaviano, & Meij, 2015)	Entity linking
(Caballero Barajas & Akella, 2015)	Naive Bayes classifier
(Oliveira, Barbar, & Soares, 2016)	multilayer perceptron, (MLP)
(Makrynioti, et al., 2017)	entity recognition
(Makrynioti, et al., 2017)	sentiment analysis

Para Dados Não estruturados

Referência	Técnica
(Sudhof, Goméz Emilsson, Maas, & Potts, 2014)	Conditional random Fields (CRFs) as a modeling technique
(Dong, et al., 2014)	Natural Language Processing
(Chen & Liu, 2014)	topic modeling
(Kurashima, Iwata, Takaya, & Sawada, 2014)	Probabilistic Latent Semantic Visualization
(Geerdink, 2015)	Data discovery
(Alshareef, Bakar, Hamdan, Abdullah, & Alweshah, 2015)	Association Rule Learning
(Makrynioti, et al., 2017)	entity recognition
(Makrynioti, et al., 2017)	sentiment analysis

Tabela 1 Técnicas para Dados Semiestruturados

Tabela 2 Técnicas para dados Não estruturados

De forma a perceber qual o valor que esta variedade de dados pode fornecer, irá ser realizado uma experimentação com o objetivo de utilizar diferentes técnicas de análise de dados, nomeadamente o reconhecimento de imagens e análise de textos através do processamento de linguagem natural. A experimentação irá recorrer à ferramenta denominada de Watson Analytics da IBM Bluemix.
Através de técnicas de reconhecimento de imagens irá ser possível obter um conjunto de descritores sobre a imagem e o tipo de taxonomia presente. Estes descritores possibilitam catalogar um conjunto de imagens com características semelhantes.
A técnica de análise de textos através do processamento de linguagem natural possibilita processar textos não estruturados e obter um conjunto de descritores, como entidades, conceitos gerais, palavras-chave, categorias, relações e análises de sentimento.

Conclusões

O trabalho realizado até ao momento possibilita perceber que a variedade de dados é um aspeto importante a ter em consideração no âmbito das organizações, pois as organizações podem estar a não aproveitar informações valiosas e que facilitem o processo de tomada de decisão.
Retirar valor de dados semiestruturados e não estruturados, obriga à perceção de técnicas analíticas existentes para essa variedade de tipos de dados, para tal foi realizada uma pesquisa em conferências científicas para identificar e perceber quais as técnicas utilizadas pela comunidade científica.
Os próximos passos consistirão em realizar várias experimentações utilizando diferentes técnicas de análise de dados do tipo semiestruturado e Não estruturado, de forma a perceber o valor que a análise deste tipo de dados pode trazer a uma organização.

Agradecimentos

This work has been supported by COMPETE: POCI-01-0145-FEDER-007043 and FCT – Fundação para a Ciência e Tecnologia within the Project Scope: UID/CEC/00319/2013.

Referências

Fan, W., & Gordon, M. D. (2014). The power of social media analytics. Communications of the ACM, 57(6), 74-81.

Russom, P. (2011). Big data analytics. TDWI best practices report, 1-35.

Uckelmann, D., Harrison, M., & Michahelles, F. (2011). An architectural approach towards the future internet of things. In Architecting the internet of things (pp. 1-24). Springer Berlin Heidelberg.

Apêndice Trabalhos Investigados

Alshareef, A. M., Bakar, A. A., Hamdan, A. R., Abdullah, S. M., & Alweshah, M. (2015). A case-based reasoning approach for pattern detection in Malaysia rainfall data. International Journal of Big Data Intelligence, 285-302.

Anagnostopoulos, C., & Triantafillou, P. (2014). Scaling out big data missing value imputations: pythia vs. godzilla. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, 651-660.

Bi, B., Ma, H., Hsu, B. J., Chu, W., Wang, K., & Cho, J. (2015). Learning to recommend related entities to search users. In Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, pp. 139-148.

Blanco, R., Ottaviano, G., & Meij, E. (2015). Fast and space-efficient entity linking for queries. In Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, pp. 179-188.

Caballero Barajas, K. L., & Akella, R. (2015). Dynamically Modeling Patient's Health State from Electronic Medical Records: A Time Series Approach. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 69-78.

Chen, N., Hoi, S. C., Li, S., & Xiao, X. (2015). SimApp: A framework for detecting similar mobile applications by online kernel learning. In Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, pp. 305-314.

Chen, Z., & Liu, B. (2014). Mining topics in documents: standing on the shoulders of big data. . In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, 1116-1125.

Dong, X., Gabrilovich, E., Heitz, G., Horn, W., Lao, N., Murphy, K., & Zhang, W. (2014). Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, 601-610.

Geerdink, B. (2015). A reference architecture for big data solutions-introducing a model to perform predictive analytics using big data technology. . International Journal of Big Data Intelligence, 236-249.

Kurashima, T., Iwata, T., Takaya, N., & Sawada, H. (2014). Probabilistic latent network visualization: inferring and embedding diffusion networks. . In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, 1236-1245.

Li, L., Yao, Y., Tang, J., Fan, W., & Tong, H. (2016). QUINT: On Query-Specific Optimal Networks.

Liu, J., Aggarwal, C., & Han, J. (2015). On integrating network and community discovery. In Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, pp. 117-126.

Makrynioti, N., Grivas, A., Sardianos, C., Tsirakis, N., Varlamis, I., Vassalos, V., & Tsantilas, P. (2017). PaloPro: a platform for knowledge extraction from big social data and the news. International Journal of Big Data Intelligence, 3-22.

Oliveira, T. P., Barbar, J. S., & Soares, A. S. (2016). Computer network traffic prediction: a comparison between traditional and deep learning neural networks. International Journal of Big Data Intelligence, 28-37.

Sudhof, M., Goméz Emilsson, A., Maas, A. L., & Potts, C. (2014). Sentiment expression conditioned by affective transitions and social force. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, 1136-1145.

Tran, N. K., Ceroni, A., Kanhabua, N., & Niederée, C. (2015). Back to the past: Supporting interpretations of forgotten stories by time-aware re-contextualization. In Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, 339-348.

Ulanova, L., Yan, T., Chen, H., Jiang, G., Keogh, E., & Zhang, K. (2015). Efficient Long-Term Degradation Profiling in Time Series for Complex Physical Systems. . In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2167-2176.

Wlodarczyk, T. W., & Hacker, T. J. (2014). Current trends in predictive analytics of big data. International Journal of Big Data Intelligence, 172-180.

Zhao, Z., Liu, J., & Cox, J. (2014). Safe and efficient screening for sparse support vector machine. . In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, 542-551.

Zhou, Y., Liu, L., & Buttler, D. (2015). Integrating vertex-centric clustering with edge-centric clustering for meta path graph analysis. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery a